Т Е Х Н И Ч Е С К И У Н И В Е Р С И Т Е Т – В А Р Н А

Гинка Калева Маринова

ИЗСЛЕДВАНЕ НА МЕТОДИ ЗА РАЗПОЗНАВАНЕ НА ДИСТРЕС ОТ ПСИХОЛОГИЧЕСКИ ТЕСТОВЕ ЧРЕЗ ИЗПОЛЗВАНЕ НА МАШИННО ОБУЧЕНИЕ

А В Т О Р Е Ф Е Р А Т

на дисертация за получаване на образователната и научна степен „ДОКТОР”

по докторска програма: Автоматизирани системи за обработка на информация и управление

професионално направление: 5.3. Комуникационна и компютърна техника

Научни ръководители: доц. д-р инж. НЕДЯЛКО НИКОЛОВ
проф. д-р инж. ТОДОР ГАНЧЕВ

Рецензенти:

1.

2.

Варна, 2022 г.

Дисертационният труд е обсъден на 08.02.2022г. в катедра „СИТ“ на катедрен съвет, съгласно заповед на Ректора на ТУ-Варна № .... /..... г. и насочен за защита.

Автор: Гинка Калева Маринова

Заглавие: Изследване на методи за разпознаване на дистрес от психологически тестове чрез използване на машинно обучение

Т Е Х Н И Ч Е С К И У Н И В Е Р С И Т Е Т – В А Р Н А

Гинка Калева Маринова

ИЗСЛЕДВАНЕ НА МЕТОДИ ЗА РАЗПОЗНАВАНЕ НА ДИСТРЕС ОТ ПСИХОЛОГИЧЕСКИ ТЕСТОВЕ ЧРЕЗ ИЗПОЛЗВАНЕ НА МАШИННО ОБУЧЕНИЕ

А В Т О Р Е Ф Е Р А Т

на дисертация за получаване на образователната и научна степен „ДОКТОР”

Варна, 2022 г.

Дисертационният труд съдържа 136 страници, включително 37 фигури, 22 таблици, 15 формули, оформени в 4 глави, приноси на дисертационния труд, списък с публикациите на автора по темата на дисертационния труд и списък на използваната литература от 144 заглавия, от които 26 на кирилица и 118 на латиница.

Защитата на дисертационния труд ще се състои на ..... г. от ..... ч. в ..... на открито заседание на жури сформирано със заповед на Ректора № ...../..... г.

Материалите по защитата (дисертацията, рецензиите и становищата) са на разположение на интересуващите се в Докторантски център, стая 318 НУК.

ХАРАКТЕРИСТИКА НА ДИСЕРТАЦИОННИЯ ТРУД

1. Актуалност на проблема

Психичното здраве на пациентите с онкологично заболяване е важно, а също така в зависимост от много фактори може да се променя във времето. Актуалността на основния проблем в дисертационния труд се определя от следните фактори:

- Динамичното развитие на причините за възникването и разпространението на психосоматични заболявания в резултат на психосоциален стрес, дистрес от драматични жизнени събития, промяна в социалния статус, урбанизация, проблеми с работата и други, са от съществено значение за здравословния психичен статус на пациента и са свързани с преодоляването на психични прояви като тревожност, стрес, дистрес, депресия и трудности с адаптацията.

- Медицинската клинична диагностика на психосоматични заболявания включва създаване, обработка и използване на значителни обеми информация за пациентите. Съвременните информационните технологии предоставят нови възможности за съхранение, обработка и извличане на нови данни, за разработка, изследване и прилагане на нови методи за диагностика и лечение.

- Методите и алгоритмите в машинното обучение са ново средство в процеса на идентификация, диагностика и оценяване на риска на заболяването. Тези инструменти, пряко или косвено могат да бъдат в помощ на лекуващия лекар, свързани с вземането на адекватни и обосновани решения на базата на натрупан личен опит и резултатите от обработката, и изследването на медицинските данни.

2. Цел и задачи на изследването

Целта на настоящия дисертационен труд е разработка на технологичен подход за синтез на характеристични описатели за диагностициране на високи нива на дистрес, изследване и оценка на основните метрики за ефективност на класификационните модели, реализирани посредством програмните платформи MATLAB и WEKA, последователно редуциране състава на описателите и свеждането им до приемлив за медицинската практика състав.

Реализацията на основната цел изисква поетапна теоретична разработка и практическо изпълнение на следните задачи:

1. Изследване на отделните признаци от скрининг за дистрес и оценка на създадените модели чрез избрани метрики за определяне на качеството на класификация.
2. Поетапна предварителна обработка на данните за извличане на зависимости и изследване на разпределението на пациентите по степените на дистрес. Определяне на корелационни зависимости за анализ и извличане на фактори, които са обект на изследване при синтез на характеристични описатели за установяване на дистрес.
3. Създаване на класификационни модели на база разработени описатели, избор на метрики за оценка на качеството и сравнение на моделите.
4. Валидация на разработените модели на характеристични описатели и диференцирано изследване на пациентите по пол и възраст.

3. Обект и предмет на изследване

Обект на изследване е приложимостта на традиционни и нови методи за синтез на характеристични описатели за бинарни данни, в комбинация с алгоритми за машинното обучение, насочени към оценка на възможностите за откриване на високи нива на дистрес**.**

Предмет на изследване са методите за синтез на характеристични описатели получени от клинични данни, съставляващи самооценка на множество индикатори от диагностицирани пациенти с дистрес, и възможностите за разпознаване на високи нива на дистрес.

4. Методи на изследване

Разработката на дисертационния труд се базира на последователното и поетапно приложение на различни алгоритми и методи за целенасочена обработка на клинични данни както следва:

- Предварителна обработка на данните посредством статистически и аналитични методи за анализ на данни и извличане на знания, базирани на сортиране, честотни изследвания и линейна корелация със средствата на пакета Microsoft Office.
- Използване на алгоритмите и методите на машинното обучение за изследване и редуциране броя на синтезираните на характеристични описатели.

5. Място на изследване

Изследванията и обработката на клиничните данни са проведени в лабораторната база на катедра СИТ при ТУ – Варна.

6. Научна новост на изследването

Разработен е технологичен подход за поетапна обработка, синтезиране на характеристични описатели и редуциране на техния състав. Разработената методика е приложима при изследването на клинични данни с широк спектър на диагностични признаци. В резултат са изследвани различни групирания на признаците и е оценена тяхната приложимост за целите на медицинската практика.

7. Практическа ценност на изследването

Практическата приложимост на получените от изследванията резултати се определя:

- Използвани са реални клинични данни на пациенти с поставена диагноза на различни нива на дистрес,
- Синтезирани са различни по състав характеристични описатели, които могат да бъдат използвани като допълнителна или съпътстваща информация при диагностициране и лечение на пациенти с дистрес.
- Използването на методиката и технологичните средства може да повиши ефективността на диагностицирането при изследвания на пациенти.

8. Апробация на изследването

Основните етапи от разработването на теоретични и приложни резултати на дисертационния труд са докладвани и публикувани в следните научни форуми и издания:

Конференции:

- 1 доклад на VIII International Scientific Conference “Engineering. Technologies. Education. Safety”, 08-11.06.2020 , Borovets;
- 1 доклад на XXIX International Scientific Conference Electronics (ET). IEEE, 2020, September 16 - 18, 2020, Sozopol, Bulgaria, **Scopus**;
- 1 доклад на International Conference on Biomedical Innovations and Applications – BIA, 24-27 Sept., 2020, Varna, Bulgaria, **Scopus**;

Списания:

- 1 статия в Годишник на Технически университет - Варна, 4(2)/2020, стр. 130-137;
- 2 статии в списанието на Компютърни науки и технологии, Година XVIII, Брой 1/2020, стр. 133-140, стр. 141-146;

Конференциите „XXIX International Scientific Conference Electronics (ET).“ и „International Conference on Biomedical Innovations and Applications – BIA“ са индексирани в международната научна база от данни “SCOPUS”.

9. Публикации по дисертационния труд

Основните етапи от разработването на дисертационния труд са отразени в 6 публикации, списък на които е приложен в края на автореферата.

СЪДЪРЖАНИЕ НА ДИСЕРТАЦИОННИЯ ТРУД

ГЛАВА 1. ОБЗОР И АНАЛИЗ НА ПРОБЛЕМИТЕ, СВЪРЗАНИ С ДИАГНОСТИЦИРАНЕТО НА ДИСТРЕС

В Първа глава е направен обзор на литературни източници, базирани на теоретични разработки и клинична практика, свързани с откриване и диагностицирането на дистрес чрез скрининг анализ на данни от специализирани въпросници. Разгледани са основни методи и средства за обработка на медицински данни. Описани са различни метрики за оценка и сравнение на резултатите от изследванията, и диагностиката. Обоснована е актуалността на проблема и изискванията за регистриране и изследване на данни от клиничните изследвания, необходимостта от прилагането на съвременни теоретични достижения и експериментални средства за обработка и извличане на данни.

ИЗВОДИ КЪМ ГЛАВА ПЪРВА

От проучването и критичният анализ за състоянието и актуалността на проблемите, свързани с клиничната диагностика на пациенти с дистрес и съответно определянето на нивото на тяхното психическо състояние са направени следните изводи:

1. Проблемът за разработването и изследването на нови описатели на дистрес е актуален поради непрекъснатото увеличаване броя на пациентите с онкологични заболявания и последващия го стрес, преминаващ в тежка форма на дистрес.
2. Използваните въпросници за определяне степента на въздействие на отделните симптоми, отчитат предимно отделни показатели и тяхното групиране не винаги дава пълноценен резултат.
3. Използваните математически средства за обработка на клиничните данни в разгледаните медицински изследвания са предимно с конкретна насоченост по отношение на определена област или заболяване и се приемат с висока степен на достоверност, т.к. се базират на добре известна математическа теория, основаваща се на разработването и доказването на хипотези.
4. Машинното обучение като съвременен подход в научните изследвания, предоставя възможност да бъдат обхванати и обработени голямо количество от данни, това е условие за синтез на характеристични описатели за дистрес, като процесът за селекция на модели е фактор за адекватността и повишаване качеството на резултатите.

ГЛАВА 2. МЕТОДИ И АЛГОРИТМИ ЗА КЛАСИФИКАЦИЯ НА ДИСТРЕС

Във Глава 2 е представен първия етап от технологичния подход за синтез, изследване на характеристичните описатели за диагностицирането на дистрес. Структурата на изследванията е базирана на следната последователност:

- Общо описание на клиничните данни, свързано с определяне на основните параметри.
- Разпределение на признаците в качеството на първоначална оценка на тяхното влияние върху нивото на дистрес.
- Изчисляване на корелация между самите признаци с цел установяване на връзка между тях, която се базира на едни и същи показатели.
- Сравнителен анализ на методи и алгоритми за класификация с цел оценка на приложимостта им при откриване на дистрес.
- Анализ на резултати от получени експериментални изследвания.

2.2. Описание и селекция на данните от клиничните изследвания

Данните от клиничните изследвания са предоставени от СБАЛОЗ “Д-р Марко Марков – гр. Варна”. Предоставената извадка съдържа 9240 записа за диагностицирани пациенти с раково заболяване с различни нива на дистрес, обхваща периода от 2016 г. до 2019 година.

2.3. Общо описание на клиничните данни от установен дистрес

Анализират се данните с цел определяне на разпределението на база възраст, ниво на дистрес и активност на признаците, регистрирани в резултат от самооценка на пациентите.

2.4. Предварителен корелационен анализ на данните

Изчислени са коефициентите на линейна корелация Kca между признаците спрямо нивото на дистрес. Изчислените коефициенти на корелация са показани в табл. 2.4 .

Таблица 2.4 *Коефициенти на корелация –дистрес/признак*

Признаци	**П6**	**П4**	**П5**	**Е3**	**П3**	**Д2**	**Ф7**	**Е1**	**Д1**	**П2**
K_{ca}	**0.76**	**0.44**	0.35	0.34	0.31	0.29	0.26	0.25	0.23	0.23
Признаци	**Е2**	**Ф1**	**Е6**	**С4**	**Е5**	**Е8**	**Ф19**	**Ф2**	**Е7**	**Ф6**
K_{ca}	0.22	0.21	0.21	0.21	0.21	0.20	0.20	0.18	0.18	0.18
Признаци	**Е4**	**Ф16**	**Ф11**	**П1**	**Ф21**	**Ф5**	**Ф10**	**С1**	**Ф4**	**Ф3**
K_{ca}	0.17	0.17	0.16	0.16	0.16	0.14	0.14	0.14	0.14	0.14
Признаци	**Ф13**	**Ф8**	**Ф23**	**Ф14**	**С3**	**Ф18**	**Ф17**	**Ф15**	**С2**	**Ф12**
K_{ca}	0.13	0.13	0.12	0.12	0.11	0.11	0.11	0.10	0.09	0.08
Признаци	**Ф9**	**Ф20**	**Ф22**							
K_{ca}	0.07	0.06	0.06							

От таблица 2.4 могат да се направят следните изводи:

- За признак П6 - Решение за лечение, е с висока корелация. Останалите признаци са със слаба зависимост.
- Поради слабата линейна корелация в процеса на диагностициране не е целесъобразно да се използват самостоятелно отделни признаци, т.к. това не води до достоверни резултати;.

Признаците от Скрининг анализа са разпределени в пет групи. Изчислени са корелационните коефициенти за всяка от групите. Получените стойности са представени в табл. 2.5.

Таблица 2.5 Корелационни коефициенти на признаците разпределени по групи

Групи за самооценка при дистрес	***Корелационен коефициент***
Общо за всички групи	0.51
Практически проблеми	**0.74**
Семейни проблеми	0.27
Емоционални проблеми	0.41
Физически проблеми	0.33
Духовни/религиозни проблеми	0.25

Получените стойности показват наличие на слаба корелация между група Семейни проблеми и ниво на дистрес и между Духовни/религиозни проблеми и ниво на дистрес. Умерена корелация се установява при групи Емоционални проблеми и Физически проблеми. Висока корелация (0.74) се наблюдава при група Практически проблеми и може да се обясни с факта, че данните се отнасят до важни, жизнени въпроси, свързани с грижата за деца, домакинство, финансови проблеми, транспорт, проблеми с работата, респ. училището, и не на последно място с решението за провеждане на лечение. Поради което налага изчисляване на коефициента на корелация, обхващащ и петте групи. Получената стойност показва наличие на значителна корелация.

ИЗВОДИ КЪМ ГЛАВА ВТОРА

От направените изследвания могат да се формулират следните изводи:

- Разпределението по възраст на пациенти с установено ниво на дистрес има ясно изразен пик в пенсионна и след пенсионна възраст, което съответства на медицинските наблюдения. Този факт се приема като обект на изследване посредством машинното обучение и съответно за синтезиране на характеристични описатели на дистрес;
- Количественото разпределение по нивата на дистрес не е равномерно. Пациентите приоритетно посочват ниски нива на дистрес. Отчитайки субективният характер на самооценката е целесъобразно изследванията да продължат при отчитане както на ниско така и високо ниво на дистрес;
- Резултатите от изследванията на „активността“ /броя на положителните отговори/ показват, че съществуват признаци. на които пациентите отговарят най-често. В последствие този факт е от съществено значение при синтезирането на описателите базирани на активността на признаците;
- Стойностите на изчислените корелационни коефициенти между броя на потвърдените отговори на пациентите и определеното от тях ниво на дистрес са ниски и не могат да се използват самостоятелно. Корелацията между групираните признаци в Скрининг анализа и нивата на дистрес е слаба. Следователно самостоятелно или групираните признаци не могат да представляват база за характеристични описатели.

ГЛАВА 3. РАЗРАБОТКА И ИЗСЛЕДВАНЕ НА ХАРАКТЕРИСТИЧНИ ОПИСАТЕЛИ ЗА ДИСТРЕС

В Глава 3 на дисертационния труд е представен втория етап от технологичния подход за синтез и изследване на характеристични описатели. В настоящата глава са изследвани различни методи за синтез на описатели обхващащи различни видове и различен брой признаци. Поставената цел е редудуциране на възможните комбинации от признаци и тяхното последователно групиране на базата на резултатите от предварителната обработка.

3.2. Постановка на проблема за синтез на характеристични описатели

Наличните набори от данни, използвани за оценка при клиничната психологическа диагностика, създават предпоставки за прилагане на съвременно лечение с подкрепа на интелигентни технологични инструменти. Сред тях са инструменти, които включват методи и алгоритми за получаване на знания, анализ на данните от медицинска гледна точка, обобщаване на полезната информация, създаване на психо-диагностични инструменти, разработване на нови видове експерименти и методи за работа с психологическа информация, базирани на съвременни компютърни технологии.

Поставянето на диагноза и определяне на нивото на дистрес се установява чрез провеждане на клинични тестове и скрининг анализ на отговори от въпросници. Въпросниците се използват по време на лечение на пациенти с онкологични заболявания. Клиничните тестове се правят с цел да се прецизира динамиката и настъпилите характеристични промени в зависимост от резултатите на проведеното лечение.

Психологическите изследвания се базират на конкретни признаци, групирани в пет групи. Посредством тях се оценява психичното състояние на пациента. Правилната диагноза и проследяването на лечението, са от съществено значение за постигне на добри и обосновани резултати за оценка, и контрол на дистрес. В настоящата глава от дисертационния труд се синтезират характеристични описатели и се изследват различни методи за предсказване на дистрес, като се използват данни получени от „*Скрининг инструмент за самооценка на дистрес*“ [81, 82, 83].

3.3. Синтез на описатели посредством групиране на признаци

Синтезът на характеристични описатели изисква селектиране и групиране на признаците, включени в скрининг анализа на пациентите. За да се извлече информативното съдържание на данните, се извършва предварителна обработка с цел получаване на допълнителна информация, необходима за определяне състава на описателите. За целите на синтез и изследване на описатели се реализира:

1. Групиране на признаците в съответствие с експертна оценка на лекарите–специалисти в следните варианти:
   - Задаване на определени тегловни коефициенти за избрани признаци от проведения „*Скрининг инструмент за самооценка на дистрес*”;
   - Селектиране признаци да се преобразуват в разрядно число по степента на двойката.
2. Групиране според резултатите от предварителната обработка на данните на базата:
   - Активност на всеки признак в предоставените данни, като е изчислена сума от положителните отговори за всички записи;
   - Сумиране на отговорите на признаци за всеки запис поотделно и изчисляване на корелационна зависимост със съответното ниво на дистрес;
3. Селектиране по групи определени от скрининг инструмента за оценка нивото на дистрес.

3.3.1. Групиране според експертната оценка

В табл. 3.2.а и табл. 3.2.б са представени признаците участващи при синтез на характеристичните описатели. Във всяка колона в табл. 3.2.а за съответния описател е записана стойност, ако същият включва признака в състава си, като за описатели от ***KR2*** до ***KR7*** са зададени тегловни коефициенти на участващия признак.

Таблица 3.1а *Представяне на признаците за разработените описатели*

Описател	П1	П2	П3	П4	П5	П6	Е1	Е2	Е3	Е4	Е5	Е6	Ф1	Ф2	Ф3	Ф4	Ф5	Ф6	Ф7	Ф13	Ф16	Ф17	Ф18	Ф19	Ф20	Ф21	Ф22	Ф23	Д1	Д2
KR1	1	1	1			1	1	1	1	1	1	1																		
KR2	1	1	2			1	2	3	2	1	1	2	3																	
KR3	1	2	1			2	2	2	1	3	3	1	2																	
KR4			1				1		1	1		1							1		1			1		1		1		
KR5		1	1				1		1	1	1	1		1			1	1	1					1		1		1		
KR6		1	1				1	1	1	1	1	1		1			1	1		1			1	1			1		1		
KR7		2	2				2	3	2	2	2	3		1			1	1		1			1	1			1		1		
KR8 *																														
KR9 *																														
KR10							1		1	1		1							1		1			1				1		
KR11							1		1	1		1							1		1			1		1		1		
KR12			1				1		1	1		1							1		1			1		1		1		
KR13			1				1		1	1		1					1		1		1			1		1		1		
KR14			1				1	1	1	1		1					1		1		1			1		1		1		
KR15		1	1				1	1	1	1		1					1		1		1			1		1		1		
KR16		1	1				1	1	1	1		1				1	1		1		1			1		1		1		
KR17		1	1				1	1	1	1		1	1			1	1		1		1			1		1		1		
KR18		1	1				1	1	1	1	1	1	1			1	1		1		1			1		1		1		
KR19		1	1				1	1	1	1	1	1	1			1	1	1	1		1			1		1		1		
KR20		1	1				1	1	1	1	1	1	1			1	1	1	1	1	1			1		1		1		
KR21				1		1																								
KR22				1	1	1																								
KR23				1	1	1			1																					
KR24			1	1	1	1			1																					
KR25			1	1	1	1			1																					1
KR26			1	1	1	1			1										1											1
KP27			1	1	1	1			1										1										1	1

В табл. 3.2.б са записани поредния степенен показател на признака включен в съответния характеристичен описател.

Таблица 3.2 б *Представяне на признаците за разработените описатели*

Описател	П1	П2	П3	П4	П5	П6	С1	С2	С3	С4	Е1	Е2	Е3	Е4	Е5	Е6	Е7	Е8	Ф7	Ф16	Ф21	Ф19	Ф23	Е1	Е2	Е3	Е4	Е5	Е6	
KR8	0	1	2	3	4	5	0	1	2	3	0	1	2	3	4	5	6	7												
KR9	0	1	2	3	4	5	0	1	2	3	0	1	2	3	4	5	6	7	0	1	2	3	4							

Разработени са пет категории характеристични описатели (категория I, II, III, IV, V), обозначени като ***KRx***. В резултат са синтезирани девет набора от характеристики на вектори, които се основават на различни подмножества на оригиналното пространство на признаците, избрани въз основа на специфични за дадена област знания и обсъдени след консултация с психо-онколог. Описателите ***KR1, KR2 и KR3*** са от категория **I** включващи 11 признаци /табл. 3.2 а./

За описател ***KR1*** са зададени тегловни коефициенти 1.

Описателят ***KR2*** е с тегловни коефициенти 1, 2 или 3. Тегловните коефициенти за описател KR3 отразяващи различно експертно мнение са със стойности 1, 2 или 3.

При съставянето на характеристичен описател категория ***II*** обозначен, като ***KR4***, е избрана подгрупа от десет признака, които частично се припокриват с признаците, избрани от категория **I**. При този описател се отчитат признаци от групата /Физически проблеми/, като целта е да се изследва тяхното въздействие върху точността при класификацията.

Съставени са характеристични описатели от категория ***III*** на базата на шестнадесет експертно оценени признака. Разгледани са три възможности за структуриране на описатели от признаци (***KR5, KR6, KR7***). Описателите ***KR5*** и ***KR6*** включените различни признаци с тегловни коефициенти **1**.

При описател ***KR7*** признаците са с различни тегловни коефициенти.

Характеристичен описател ***KR8*** от категория ***IV*** е специфичен, като тегловните коефициенти на включените признаци приемат стойности по степените на двойката /табл. 3.2 б./. Осемнадесетте признака са избрани на базата на експертно мнение от психо-онколог.

За описател ***KR 9*** от категория ***V*** е от категория ***IV*** броят на признаците в подмножеството се увеличава до двадесет и три.

3.3.2 Групиране според резултатите от предварителната обработка

Въз основа на предварителната обработка на данните за група описатели са характерни три случая:

1.Описателите от ***KR10*** до ***KR20*** са определени на база на тяхната активност (броя на отговорите отбелязани с **1** от всички предоставени записи). Ако с $\boldsymbol{\beta} = \{\beta_j\}$ където $\boldsymbol{j \epsilon \{1 \ldots 43\}}$ се обозначи множеството на признаците и съответно с $A = \{\boldsymbol{a_{ij}}\}$ множество на записите, може да се изчислят сумите на отделните записи, представени с множеството $\boldsymbol{S} = \{\boldsymbol{s_i}\}$ по формулата:

$$(\forall \beta_j \, \boldsymbol{\epsilon} \, \boldsymbol{\beta})(\boldsymbol{s_j} = \sum_{i=1}^{N} \boldsymbol{a_{ij}}) \tag{3.4}$$

Получените суми са сортирани в низходящ ред по правилото:

$$(\forall \boldsymbol{s_j}, \boldsymbol{s_{j+1}} \boldsymbol{\epsilon} \, \boldsymbol{S}) \rightarrow (\boldsymbol{s_j} \geq \boldsymbol{s_{j+1}}) \tag{3.5}$$

На база сортиране се подреждат отделните признаци по реда на тяхната активност, изчислена като сума от всички записи, и се използва като критерий при определяне състава на описателите.

2.Описатели получени за всяка от петте групи от представените записи самостоятелно. Целта на това изследване е да се определи до каква степен признаците от дадена група доминират над останалите групи.

3.За получаване на характеристичните описатели от KR21 до KR27 е извършено сумиране на признаците за всеки запис в зависимост от изчислената корелация за всеки признак спрямо дистреса.

3.3.3. Селектиране по групи определени от Скрининг инструмента за оценка нивото на дистрес

Скрининг инструмент за самооценка на дистрес (табл. 1. 3) е разделен на пет групи признаци *(„Практически проблеми“, „Семейни проблеми“, „Емоционални проблеми“, „Духовни/религиозни проблеми“ и „Физически проблеми“*). Всяка от определените групи може да се разглежда, като отделен характеристичен описател.

Описател ***KR28*** включва признаците от групата „Практически проблеми“, обозначени с П1, П2, П3, П4, П5 и П6.

Описател ***KR29*** включва ***4*** признака от групата „Семейни проблеми“, обозначени със С1, С2, С3 и С4.

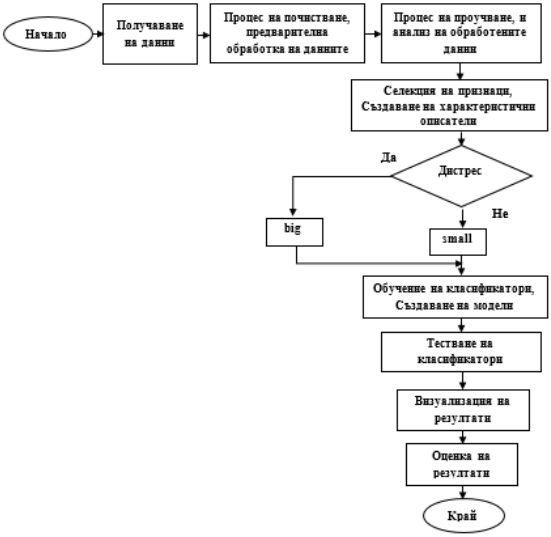
Описател ***KR30*** включва признаците от групата „Емоционални проблеми“, обозначени с Е1, Е2, Е3, Е4, Е5, Е6, Е7 и Е8.

Описател ***KR31*** включва признаците от групата „Духовни/религиозни проблеми“, и са обозначени с Д1 и Д2.

Описател ***KR32*** включва ***23*** признаци от групата „Физически проблеми“, обозначени последователно от Ф1 до Ф23 включително.

3.4 Изследване на характеристични описатели

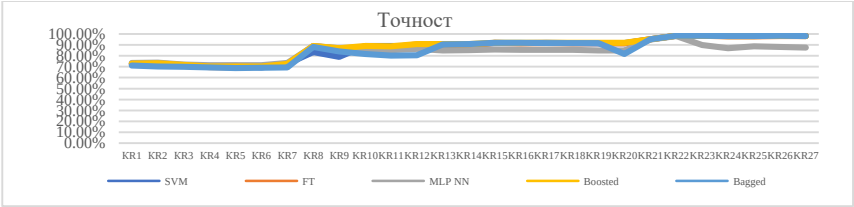
Представени са експериментални резултати и е направен сравнителен анализ на синтезираните характеристични описатели, получени чрез оценка на ефективността на следните класификатори: FT, SVM, Boosted, Bagged и MLP NN. Класификаторите са създадени чрез методите дърво на решения, опорни вектори, ансамблови алгоритми и неврона мрежа. На фиг. 3.2 е показана алгоритмичната процедура за класификация на дистрес.



Фиг. 3.2 Процес на класификация на дистрес

Експерименталните резултати за показател точност са представени на фиг. 3.3. Точността е показател за относителния брой правилно класифицирани данни в модела

спрямо общия брой обекти в извадката. Тази метрика е една от най-често използваните при оценка на класификаторите.



Фиг. 3.3. Графика на резултатите за точност на класификаторите

Получените резултати за мярката точност са в диапазона от 68.70%. до 98.4%, които могат да се оценят, като сравнително високи за създадените 135 модела. Горната граница от 98.4% се постига от моделите създадени на база описатели ***KR22*** и ***KR25***. Анализирайки получените резултати се наблюдава следната тенденция:

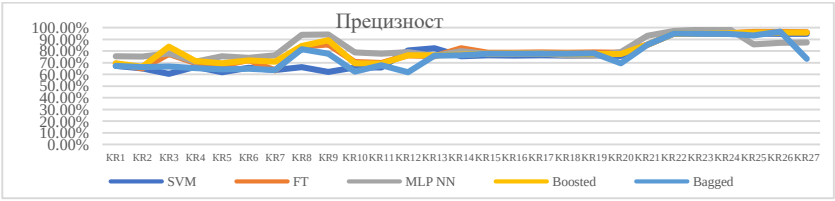
• Характеристичните описатели от ***KR1*** до ***KR7*** имат ниска точност в сравнение с останалите, което трябва да се има в предвид при тяхната комплексна преценка.

• Забелязва се значителна динамика на локално изменение при SVM, Bagged и MPL NN. Моделите за FT и Boosted реализират близки стойности за метрика точност, които са без съществени изменения при различните описатели.

При класификация на двадесет и седем характеристични описатели най-висока точност се отчита при модели създадени върху на характеристичен описател ***KR22*** (Метод дърво на решения, Ансамблов алгоритъм Boosted, Ансамблов алгоритъм Bagged) и характеристичен описател ***KR25*** (Метод на опорните вектори) -98.4%.

• Точността е най-ниска при класификатор Bagged за описател ***KR5*** 68.70%.

Частта на правилно разпознатите обекти от клас положителен спрямо общия брой обекти, приети от класификатора като обекти от клас положителен, е представена на фиг. 3.4.

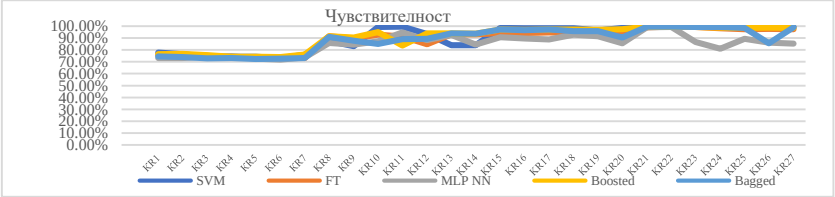


Фиг. 3.4 *Графика на резултатите за прецизност на класификаторите*

На фиг.3.4 се забелязва значителна динамика на изменение в стойностите на изчислената прецизност на класификаторите в зависимост от изследваните описатели. Висока прецизност се регистрира от класификатори MPL NN за характеристични описатели - ***KR8, KR9***, и от ***KR21*** до ***KR27***. За модели, създадени с метод на опорните вектори, (SVM) прецизността е незадоволителна за описатели от ***KR1*** до ***KR12*** (от 60.60% до 80.42%). Моделите създадени с други методи на машинно обучение за характеристични описатели ***KR8***, ***KR9,*** са с по-висока прецизност сравнение с моделите

на SVM над 15%. Прецизността на моделите създадени с четири метода на машинно обучение на характеристични описатели от ***K13*** до ***KR27*** са с близки стойности. Разликата е в порядъка около10%.

Чувствителност е мярка при изследваните модели, които правилно са идентифицирали истинските положителни резултати.

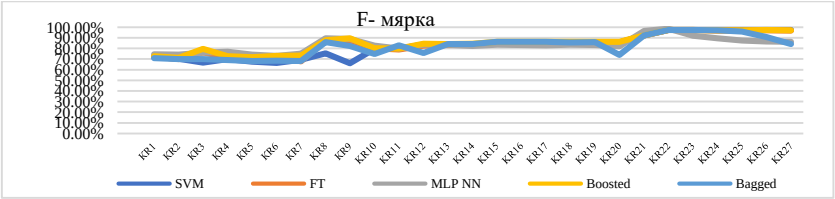


Фиг. 3.5 Графика на резултатите за чувствителност на класификаторите

Резултатите показват (фиг. 3.5), че най-висока чувствителност се получава за модел, създаден по метод на опорните вектори за описател ***KR25*** (99.8%). Чувствителността е висока за модели за петте класификатора над 99% за характеристични описатели ***KR21, KR22*** и за описатели ***KR10, KR11*** и за ***KR21 до KR27*** за модели формирани по метод на опорните вектори.

Най-ниска е чувствителността за характеристичен описател ***KR6*** при използване на MLP NN (71.69%). Чувствителността е ниска за ***KR1*** до ***KR8*** за всички изследвани класификатори.

F-мярката е общ критерий за качество, основаващ се на прецизност и чубствителност. Експерименталните резултати за тази метрика са показани в фиг. 3.6.



Фиг. 3.6 Графика на резултатите за Ф мярка на класификаторите

От резултатите представени на фиг.3.6., могат да се направят следните заключения: Описатели ***KR1 до KR8*** показват нисък процент за F-мярка. В определени зони за класификаторите SVM и Bagged може да се оцени като динамична. Получените стойности за F-мярка за FT и Boosted се препокриват в определени граници**.** Описатели ***KR21*** до ***KR23*** показват най-висок процент за показател F-мярка за всички класификатори**.**

3.6. Сравнителна оценка на разработените описатели

За синтезираните 32 характеристични описатели за метриката **точност** най-високи резултати са получени по корелация от ***KR21*** до ***KR27***, като моделите създадени за класификаторите SVM, FT, Bagged и Boosted, за описатели ***KR22*** и ***KR25*** са с резултати

над 98%, следвани от групата основаваща се на висока активност според дадените отговори от пациентите, като най-висока стойност е получена за ***KR17*** 91.90%. С ниски резултати по точност са описателите, селектирани по групи.

За **прецизност** най-добри резултати са получени за моделите за характеристични описатели по корелация. С недобро представяне са описателите селектирани по групи.

За метриката **чувствителност** за описателите по корелация от ***KR21*** до ***KR27*** и за характеристични описатели по активност ***KR10*** и ***KR11*** резултатите са най-високи над 99%. Най-нисък получен резултат е за модел на характеристичен описател ***KR6*** 71.69%.

За **F-мярка** най-добри резултати са получени за описателите по корелация. С най-висок резултат е получен за характеристичен описател ***KR22*** - 98.64%. С най-лошо представяне е характеристичен описател ***KR20*** 64.04%.

3.7. Изводи

Синтезирани и оценени са модели на тридесет и два характеристични описатели. Реализирана е класификационна задача за изследване в медицинската област психоонкология, която е свързана с обработка на данни в подкрепа на процеса на вземане на решения. Основните резултати са:

- Въведени са критерии за групиране на признаците в характеристични описатели, включващи експертните оценки на онкопсихолози, тегловни коефициенти, на базата на предварителната обработка от предната глава свързана с активността на признаците, установената корелация с нивата на дистрес и известните пет групи от скрининг анализа;
- Обработени и анализирани са данните на пациентите с поставена онкологична диагноза, като са отстранени личните данни;
- При проведеното експериментално изследване са използвани пет варианта на машинно обучение, за създаване на модели. Използваните методи и алгоритми от машинното обучение представят различни резултати според обработените данни;
- Изследвани са модели на тридесет и два характеристични описатели, като са оценени чрез метрики точност, прецизност, чувствителност и F-мярка . Извършен е сравнителен анализ. Резултатите за показател точност са в диапазона от 68.70% до 98.40.%. С най-висока стойност за прецизност е характеристичния описател ***KR24*** 99.83%. За метриката чувствителност стойностите на резултатите са в диапазона от 71.69% до 99.80%. Получените резултати за създадените модели за характеристичните описатели ***KR22*** и ***KR25*** са сравнително по-точни, което е целесъобразно да се вземе в предвид при диагностицирането и последвалото лечение.

ГЛАВА 4. ВАЛИДАЦИЯ И ИЗСЛЕДВАНЕ НА ХАРАКТЕРИСТИЧНИ ОПИСАТЕЛИ ЗА ВИСОКИ НИВА НА ДИСТРЕС

4.1. Цел и задачи за валидация на характеристични описатели

Основната цел и задачите на изследванията в Глава 4 са свързани с валидация на описателите и диференцирано изследване на пациенти, групирани по пол и възраст.

Формулирани са следните задачи:

- Разработка на алгоритъм за съставяне на валидираща таблица и сравнителен анализ на резултатите;
- Повторно изследване на разработените описатели и тяхното валидиране със средствата на платформата WEKA;
- Оценка на характеристичните описатели и редуциране на техния брой, без да се намалява общата им ефективност;
- Разпределение на данните по пол на пациентите, изследване на описателите и преценка на тяхната ефективност и приложимост;
- Разпределение на пациентите по възраст и повторно изследване на предложените описатели.

4.2. Същност на валидацията на описатели

За постигане на целта са представени алгоритъм и валидираща таблица за определените характеристични описатели посредством MATLAB и WEKA. Представено е редуциране броя на характеристични описатели и тяхната валидация посредством платформата WEKA, и е извършен сравнителен анализ на резултатите от MATLAB и WEKA. Проведено е изследване на редуцираните описателите при селектиране на записите по пол и възраст на пациентите.

4.3. Разработка на алгоритъм за съставяне на валидираща таблица

Намален е броят на характеристичните описатели до 12, като същите са изследвани по отношение на метриките точност, прецизност, чувствителност и F-мярка. Като се има в предвид значимостта на получените до момента резултати по отношение диагностиката на високи нива на дистрес и субективния характер на основните данни, е необходима тяхната повторна валидация.

За изпълнението на това изискване и за постигането на по-висока сигурност на резултатите се предлага алгоритъм за последващо редуциране броя на характеристичните описатели. Основните компоненти на алгоритъма са:

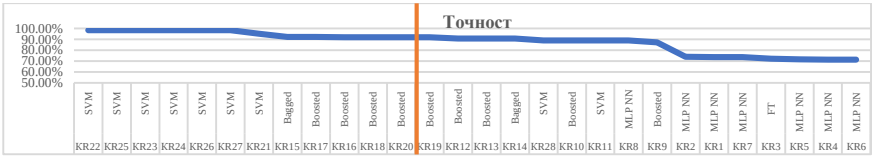
- Изследвани са определени характеристичните описатели селектирани от предната глава. Точност на класификация е основна метрика за оценка. Описателите са сортирани по този показател и е определена граничната стойност;
- Характеристичните описатели са редуцирани на база определената гранична стойност;
- Останалите метрики прецизност, чувствителност и F-мярка са приведени към точността по отношение на редуцираните описатели и използваните методи за обучение, т.е. условията на изследване са напълно идентични;
- За резултатите получени с MATLAB и WEKA по двойка метрика точност, прецизност, чувствителност и F-мярка, е изчислена тяхната линейна корелация;
- Определя се диапазона въз основа на минималната и максимална стойност на съответната метрика;
- Посредством получените резултати и направен извод за оценка на валидацията на данните.

4.4. Редуциране на характеристични описатели по критерий за точност

4.4.1. Критерии за избор и редуциране на характеристичните описатели

В Глава 3 са разработени и изследвани 27 характеристични описатели с близки по стойност показатели като точност, чувствителност, прецизност и F-мярка.

За метриката **точност** описателите са сортирани в низходящ ред /фиг.4.3/. В диапазона от 98.40% до 92% се намират 12 характеристични описатели. Точността на следващите е значително по-ниска и поради тази причина тяхното валидиране и приложение на този етап е нецелесъобразно.



Фиг. 4.3. Сортиране на описателите по мярката точност

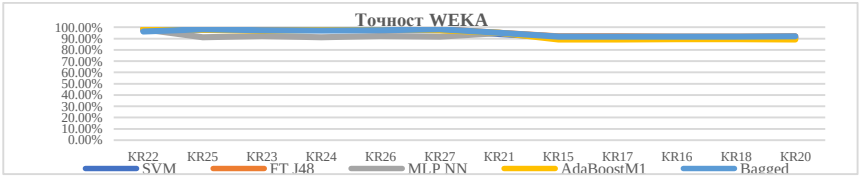
4.4.2. Изследване на редуцираните описатели посредством WEKA

Постановката на повторните изследвания на определените характеристични описатели е следната:

- броят и състава на описателите не е променен;
- подредбата на характеристичните описатели е запазена;
- във WEKA са използвани същите методи за машинно обучение, който са използвани в MATLAB;
- резултатите са регистрирани таблично и са визуализирани графично при едни и същи метрики за оценка на класификацията.

Получените резултати от реализираното изследване потвърждават ефективността на дванадесетте характеристични описатели, а също така е осъществена експериментална оценка на класификационната точност.

На фиг. 4.7 е представена сравнителна графика за оценката на класификаторите по метриката точност.

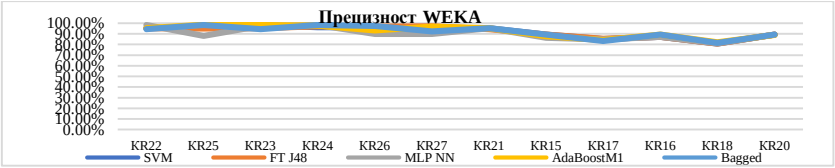


Фиг. 4.7 Графика на Точност получена с WEKA

Точността на класификационните модели е в диапазона от 89.06%. до 98.96%. Класификатор AdaBoostM1 за характеристичен описател KR22 реализира най-висока точност 98.96%. На второ място е класификатор FT J48 с точност 98.59%. за характеристичен описател KR22. На трето място е KR22 за класификатор Bagged 96.35%. Получените резултати потвърждават предварителното предположение за проверка на описателите и тяхната ефективност. За характеристичните описатели, базирани на

корелация (***KR22, KR25, KR23, KR24, KR26, KR27***), точността е в диапазона от 98.96% до 91.84% (разликата е 7.12%) това е една висока точност на класификация. За описателите, определени на базата на положителните отговори, точността е в диапазона от 95.10% до 89.06% (разликата е 6.04%). В заключение може да се обобщи, че характеристичните описатели, базиращи се на корелация, могат да бъдат използвани успешно в задачи насочени към класификацията на дистрес.

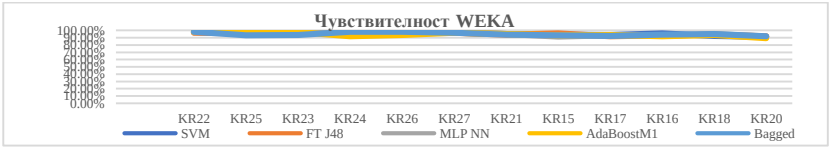
За създадените 60 модела с пет метода на машинно обучение е изчислена метриката прецизност. Получените резултати са представени в на фиг. 4.8.



Фиг. 4.8 Графика на Прецизност получена с WEKA

Анализът на резултатите, на получените стойностите за прецизност за изследваните описатели е сравнително висока. Максимално получената стойност е 98.70%, а минималната 80.72%. Класификаторите създадени чрез алгоритъм AdaBoostM1 реализират най-високи стойности за метрика прецизност, следвани от Bagged и SVM.

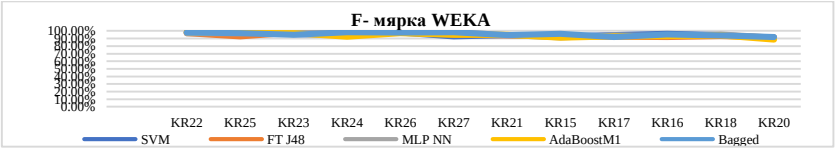
На фигура 4.9 са представени изчислените стойности за метрика чувствителност.



Фиг. 4.9 Графика на Чувствителност получена с WEKA

На фигура 4.9 са визуализирани резултатите за мярката чувствителност за отделните характеристични описатели представени в графична диаграма. Оценката е на базата на сравнителен анализ, като се установява, че максимално получената стойност е **98.20%,** а минималната е **89.30%** за характеристичен описател ***KR20***, за класификатор AdaBoostM1.

Създадените 60 класификатора са съпоставени с F-мярка. Получените резултати са показани на фигура 4.10.



Фиг. 4.10 Графика на F-мярка получена с WEKA

Максималната стойност на получената оценка за **F-мярка** е **98.82%** за характеристичен описател ***KR25***, за SVM и FT J48, Bagged.

4.5. Разработка на валидираща таблица

Таблица 4.5 е разработена на следния принцип:

- За всяка метрика точност, прецизност, чувствителност и F-мярка данните са подредени самостоятелно по двойки на базата на резултатите от MATLAB и WEKA;
- Последователността на характеристичните описатели е запазена, като и използваните методи за машинно обучение;
- Изчислена е корелацията, границите и диапазона на изменение.

Таблица 4.5 Валидираща таблица на основните метрики получени посредством MATLAB и WEKA

Мярка	Платформа	KR22	KR25	KR23	KR24	KR26	KR27	KR21	KR15	KR17	KR16	KR18	KR20	Корелация	min	max	Разлика
Точност	MATLAB	98.40%	98.40%	98.30%	98.30%	98.30%	98.30%	95.30%	92.10%	92.10%	92.00%	92.00%	92.00%	**0.98**	92.00%	98.40%	6.40%
	WEKA	97.24%	98.24%	98.02%	98.11%	97.59%	98.15%	94.08%	91.82%	89.06%	89.58%	89.42%	89.27%		89.06%	98.24%	9.18%
Прецизност	MATLAB	95.10%	95.12%	94.85%	94.85%	94.85%	94.85%	85.45%	77.71%	77.86%	77.98%	77.68%	77.52%	**0.88**	77.52%	95.12%	17.60%
	WEKA	95.30%	96.30%	98.20%	96.20%	96.30%	95.42%	95.30%	89.22%	84.26%	89.23%	82.17%	89.06%		82.17%	98.20%	16.03%
Чувствителност	MATLAB	99.60%	99.80%	99.45%	99.30%	99.00%	99.25%	99.20%	97.15%	96.96%	96.78%	97.19%	97.23%	**0.81**	96.78%	99.80%	3.02%
	WEKA	98.20%	97.82%	98.10%	93.81%	95.82%	96.81%	94.90%	93.10%	93.61%	92.10%	93.88%	89.30%		89.30%	98.20%	8.90%
Ф мярка	MATLAB	97.49%	97.50%	97.36%	97.36%	97.36%	97.36%	92.15%	86.35%	86.37%	86.37%	86.35%	86.26%	**0.59**	86.26%	97.50%	11.24%
	WEKA	98.20%	98.82%	95.80%	94.81%	97.82%	92.81%	94.60%	96.17%	92.84%	93.40%	94.20%	88.60%		88.60%	98.82%	10.22%
Класификатор		SVM	SVM	SVM	SVM	SVM	SVM	SVM	Bagged	Boosted	Boosted	Boosted	Boosted				

Фиг. 4.11 Графики на метрики получени посредством MATLAB и WEKA

4.6. Обобщение, анализ и изводи от валидацията на описателите

От получените резултати могат да се направят следните изводи:

- За мярката **Точност** диапазоните между минималната и максимална стойност са сравнително малки и до известна степен се препокриват;
- Коефициентът на корелация за точността между резултатите, получени от MATLAB и WEKA е много висока със стойност 0.9844. Това дава основание да се приеме, че използването на програмните средства дават приблизително еднакви резултати.
- MATLAB реализира по-висока точност при класификация в сравнение с WEKA (фиг.4.11);
- От графика 4.11 се наблюдава, че точността, намалява след шестия описател, но запазва високи стойности;
- От сравнителния анализ на резултатите на дванадесетте изследвани описатели (табл.4.5) се установява, че разликата между минималната и максимална стойност за мярката **Прецизност** е незначително увеличена в сравнение с тази на точността;

- Коефициентът на корелация е 0.8837, което потвърждава предположението за близост на резултатите и съответно на тяхното изменение при използването на различни програмни средства, реализиращи машинно обучение;
- След седмия описател прецизността изчислена с WEKA, е по-висока в границите на определения диапазон;
- При метрика **Чувствителност,** резултатите получени с MATLAB определено са по-високи и диапазона на изменение е 3.02%, докато при WEKA те са по-ниски.
- Коефициентът на корелация, отчитащ основно свързаността и изменението на резултатите при изчисленията на двете програми, се запазва сравнително високи 0.8069;
- При **F-мярката** изменението на резултатите е динамично и съответно коефициента на корелация между тях е 0.5920;
- Диапазонът на изменение може да се оцени като значителен;

4.7. Изследване на характеристичните описатели на база пол на пациент

4.7.1. Постановка, предпоставки и цел на изследване

Обект на изследване са клинични данни на пациенти с диагностицирано онкологично заболяване. Основната цел на изследването е посредством методите на машинното обучение при използване на MATLAB и WEKA да се предскаже дистрес както при мъжете, така и при жените. Създадените класификатори са съпоставени чрез метрики точност, прецизност, чувствителност и F-мярка на жените и отделно на мъжете. Поставените задачи са:

- Да се изследват описателите подредени по същия начин, според направената сортировка по точност;
- Да се използват едни и същи класификатори на машинното обучение с MATLAB и WEKA;
- Да се определи минималната и максимална граница и разликата между тях като косвена оценка за диапазона на изменение за всяка оценъчна метрика;
- Изчисляване на корелационни коефициенти;.

Изчислените метрики за създадените модели са представени в табл. 4.6 и 4.7.

Таблица 4.2 *Изследване на пациенти с дистрес - мъже*

Мярка	Платформа	KR22	KR25	KR23	KR24	KR26	KR27	KR21	KR15	KR17	KR16	KR18	KR20	Корелация	min	max	Разлика
Точност	MATLAB	98.80%	98.80%	98.80%	98.80%	98.80%	94.80%	96.20%	92.50%	92.60%	92.40%	92.60%	92.70%	**0.98**	92.40%	98.80%	6.40%
	WEKA	98.59%	98.48%	98.62%	98.67%	98.38%	93.20%	96.01%	92.64%	90.46%	90.30%	90.50%	90.56%		90.30%	98.67%	8.37%
Прецизност	MATLAB	95.85%	96.11%	94.99%	95.75%	95.07%	88.86%	87.74%	79.45%	77.98%	77.46%	77.81%	77.98%	**0.96**	77.46%	96.11%	18.65%
	WEKA	98.36%	98.50%	98.60%	98.70%	98.38%	93.41%	95.90%	93.00%	91.60%	91.50%	91.00%	92.20%		91.00%	98.70%	7.70%
Чувствителност	MATLAB	99.28%	99.82%	99.64%	98.89%	98.75%	95.73%	97.25%	94.41%	94.63%	94.39%	94.53%	94.63%	**0.97**	94.39%	99.82%	5.43%
	WEKA	98.60%	98.50%	98.60%	98.70%	98.40%	93.40%	96.20%	92.60%	90.60%	90.60%	90.60%	90.60%		90.60%	98.70%	8.10%
Ф мярка	MATLAB	97.54%	97.93%	97.26%	97.29%	96.88%	92.17%	92.25%	86.29%	85.50%	85.09%	85.36%	85.50%	**0.96**	85.09%	97.93%	12.84%
	WEKA	98.10%	98.20%	97.90%	98.00%	98.40%	93.05%	96.00%	92.40%	92.30%	90.00%	92.30%	90.20%		90.00%	98.40%	8.40%
Класификатор		SVM	SVM	SVM	SVM	SVM	SVM	SVM	Bagged	Boosted	Boosted	Boosted	Boosted				

***Таблица 4.7** Изследване на пациенти с дистрес – жени*

Мярка	Платформа	KR22	KR25	KR23	KR24	KR26	KR27	KR21	KR15	KR17	KR16	KR18	KR20	Корелация	min	max	Разлика
Точност	MATLAB	97.80%	97.70%	97.80%	97.80%	97.80%	93.20%	94.30%	91.10%	91.50%	91.50%	91.90%	91.50%	**0.91**	91.10%	97.80%	6.70%
	WEKA	97.77%	97.74%	97.74%	97.85%	97.68%	97.65%	94.25%	90.11%	90.12%	90.07%	90.78%	89.96%		89.96%	97.85%	7.89%
Прецизност	MATLAB	94.67%	95.14%	93.13%	93.12%	93.48%	93.58%	83.09%	78.19%	77.05%	77.00%	76.13%	77.43%	**0.98**	76.13%	95.14%	19.01
	WEKA	97.75%	97.85%	97.80%	97.86%	97.70%	97.70%	94.70%	90.50%	91.60%	91.50%	91.20%	90.30%		90.30%	97.86%	7.56%
Чувствителност	MATLAB	98.18%	98.79%	98.13%	97.89%	98.06%	98.09%	95.74%	93.19%	93.52%	93.58%	93.14%	93.56%	**0.99**	93.14%	98.79%	5.65%
	WEKA	97.65%	97.90%	97.70%	97.84%	97.70%	97.70%	94.20%	90.10%	90.10%	90.15%	90.80%	90.00%		90.00%	97.90%	7.90%
Ф мярка	MATLAB	96.40%	96.93%	95.56%	95.44%	95.72%	95.78%	88.97%	85.04%	84.49%	84.48%	83.78%	84.74%	**0.99**	83.78%	96.93%	13.15
	WEKA	97.78%	97.81%	97.70%	97.80%	97.70%	97.70%	94.10%	89.80%	89.70%	89.80%	88.90%	89.60%		88.90%	97.81%	8.91%
Класификатор		SVM	SVM	SVM	SVM	SVM	SVM	SVM	Bagged	Boosted	Boosted	Boosted	Boosted				

Фиг. 4.14 Графики на метрики получени посредством MATLAB и WEKA мъже

Фиг. 4.15 Графики на метрики получени посредством MATLAB и WEKA жени

4.7.2. Обобщение, анализ и изводи от проведените изследвания.

От получените резултати на проведените експериментални изследвания по пол, показани /таблица 4.6 и таблица 4.7, и фиг.4.14, и фиг. 4.15/ са направени следните изводи:

1. По отношение на метриката **Точност**:

- Коефициентите на корелация за метриката **Точност** са виоки. За мъже е **0.9843**, за жени съответно **0.9115** с разлика **0.0728**;
- **Точността** получена от MATLAB, сравнение с WEKA е малко по-висока, но тя се приема за несъществена;
- От обобщените данни се установява, че разлики между минимална и максимална стойност са малки;
- От Фиг. 4.14 за мъже и Фиг. 4.15 жени е визуализирано изменението на точността.

2. По отношение на метриката **Прецизност:**

- Коефициентите на корелация за **Прецизност** са високи, като при мъже е **0.9669** за жени е **0.9814** с разликата **0.0146**;
- Разликата между минималната и максимална стойност за мярката **Прецизност** е малко увеличена в сравнение с тази за показател **Точност**;
- Прецизността, изчислена с WEKA, е по-висока сравнение с тази на MATLAB, като след петия описател е обратно.

3. По отношение на метриката **Чувствителност:**

- Коефициентите на корелация, получени за мярката **Чувствителност** са високи, за мъже е **0.9766,** за жени са малко по-високи **0.9919**, разликата е **0.01533;**
- Диапазонът на изменение на стойностите за Чувствителност се запазва под 10 %, което е в унисон със субективния характер на изследваните данни.

4. По отношение на **F-мярката:**

- За **F-мярка** от резултатите се констатира рязък спад след седми описател, коефициентът на корелация за мъже е **0.9570**, за жени е **0.9931**, разликата между тях е **0.0361**.
- Обхватът на изменение на резултатите може да се оцени, като значителен.

4.8. Изследване на описатели на база възраст на пациент

4.8.1. Постановка, предпоставки и цел на изследването

Онкологични заболявания се регистрират във всички възрастови групи в световен мащаб. Броят на регистрираните пациенти със злокачествени заболявания за възрастните хора е с по-голям дял на заболяваемост, в сравнение с по-младите. Целта на настоящето изследване е да се установи до каква степен отделните възрастови сегменти се отразяват на основните показатели за точност, прецизност, чувствителност и F-мярка. Постановката и условията на изследванията са аналогични на предните, като се запазва поредността на описателите и използваните методи на машинно обучение. Групирането на пациентите в сегменти по възраст в изследваните източници е различно, като това зависи от целта на изследването, от специфичните особености на заболяването и от въздействието на различни външни фактори.

За целите на настоящото изследване пациентите са разделени на три групи.

I група. В нея са включени пациенти от 19 до 30 години, които представляват **0.32%** от всички изследвани пациенти с установена диагноза – дистрес. Незначителният брой пациенти включени в тази група е определен от първоначалните данни за ниска онкологична заболеваемост и последвал дистрес.

II група. В сегмента попадат пациенти на възраст от 31 до 65 г. с дял от **36.35%.** Това са пациенти в работоспособна възраст с повишена степен на онкологични заболявания в сравнение с първата група.

III група. В тази група са включени пациенти на пенсионна и след пенсионна възраст **(63.32%).** Характерното за тази група е недоброто здравословно състояние,

занижен имунитет и както показват статистическите данни с най-голям процент за онкологични заболявания.

Таблица 4.3 *Изследване на пациенти с дистрес възраст 19-30 години- I група*

Мярка	Платформа	KR22	KR25	KR23	KR24	KR26	KR27	KR21	KR15	KR17	KR16	KR18	KR20	Корелация	min	max	Разлика
Точност	MATLAB	91.18%	96.67%	82.76%	90.00%	93.33%	90.00%	86.67%	73.33%	76.67%	83.33%	86.67%	93.10%	**0.52**	83.20%	95.50%	12.30%
	WEKA	83.33%	98.58%	73.33%	80.00%	83.33%	85.19%	83.39%	73.33%	90.00%	86.67%	86.67%	90.30%		73.33%	98.58%	25.25%
Прецизност	MATLAB	75.00%	100.00	66.67%	70.00%	80.00%	88.89%	80.00%	66.67%	60.00%	60.00%	80.00%	80.00%	**0.45**	60.00%	100.00%	40.00%
	WEKA	86.70%	96.60%	73.70%	84.60%	86.70%	87.90%	86.70%	72.20%	91.20%	91.00%	91.30%	91.15%		72.20%	96.60%	24.40%
Чувствителност	MATLAB	91.34%	95.35%	85.01%	90.35%	92.82%	90.96%	88.03%	76.49%	79.85%	85.08%	88.03%	92.50%	**0.45**	76.49%	95.35%	18.86%
	WEKA	83.30%	95.70%	73.30%	80.00%	83.30%	85.20%	83.30%	73.30%	90.00%	90.40%	90.10%	90.20%		73.30%	95.70%	22.40%
Ф мярка	MATLAB	82.37%	97.62%	74.73%	78.88%	85.94%	89.91%	83.82%	71.24%	68.52%	70.37%	83.82%	85.80%	**0.43**	68.52%	97.62%	29.10%
	WEKA	81.50%	97.30%	71.40%	77.00%	81.50%	83.80%	81.50%	72.20%	89.60%	89.30%	89.35%	89.50%		71.40%	97.30%	25.90%
Класификатор		SVM	SVM	SVM	SVM	SVM	SVM	SVM	Bagged	Boosted	Boosted	Boosted	Boosted				

Фиг. 4.17 Графики на метрики получени за възраст 19-30 години- I група

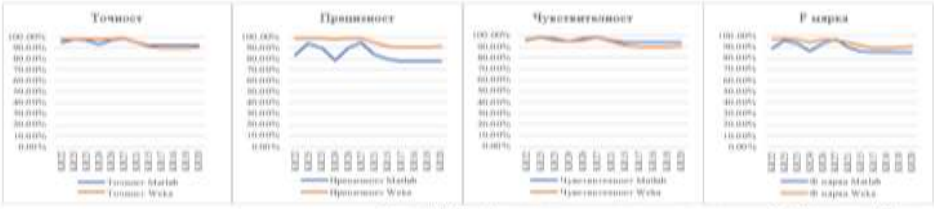
Таблица 4.5 *Изследване на пациенти с дистрес възраст 31-65 години- II група*

Мярка	Платформа	KR22	KR25	KR23	KR24	KR26	KR27	KR21	KR15	KR17	KR16	KR18	KR20	Корелация	min	max	Разлика
Точност	MATLAB	91.96%	98.48%	94.61%	93.12%	98.48%	98.52%	94.61%	91.46%	91.75%	91.90%	91.54%	92.05%	**0.86**	91.46%	98.52%	7.06%
	WEKA	93.21%	98.24%	97.07%	96.12%	97.30%	98.35%	95.56%	91.57%	89.37%	89.34%	89.70%	89.13%		89.13%	98.35%	9.21%
Прецизност	MATLAB	84.86%	95.28%	83.26%	78.63%	95.28%	95.38%	83.26%	77.98%	76.60%	80.00%	76.78%	77.52%	**0.74**	76.60%	95.38%	18.79%
	WEKA	93.30%	98.30%	97.50%	98.50%	98.20%	98.40%	95.80%	92.10%	90.60%	90.45%	90.36%	90.40%		90.36%	98.50%	8.14%
Чувствителност	MATLAB	93.58%	98.86%	96.01%	94.89%	98.86%	98.89%	96.01%	93.57%	93.82%	92.88%	93.64%	94.05%	**0.86**	92.88%	98.89%	6.01%
	WEKA	92.20%	98.20%	97.35%	98.40%	98.30%	98.50%	95.60%	91.60%	90.40%	89.10%	89.60%	90.20%		89.10%	98.50%	9.40%
Ф мярка	MATLAB	89.01%	97.04%	89.18%	86.00%	97.04%	97.10%	89.18%	85.07%	84.34%	79.48%	84.38%	84.99%	**0.78**	79.48%	97.10%	17.62%
	WEKA	91.20%	98.40%	97.25%	98.35%	98.30%	98.50%	95.00%	91.30%	90.00%	90.40%	89.60%	90.20%		89.60%	98.50%	8.90%
Класификатор		SVM	SVM	SVM	SVM	SVM	SVM	SVM	Bagged	Boosted	Boosted	Boosted	Boosted				

Фиг. 4.18 Графики на метрики получени за възраст 31-65 години- II група

Таблица 4.10 *Изследване на пациенти с дистрес възраст 66-99 години- III група*

Мярка	Платформа	KR22	KR25	KR23	KR24	KR26	KR27	KR21	KR15	KR17	KR16	KR18	KR20	Корелация	min	max	Разлика
Точност	MATLAB	94.62%	98.10%	96.48%	93.06%	96.58%	98.48%	94.84%	91.93%	92.22%	92.00%	91.97%	92.19%	**0.82**	91.93%	98.48%	6.54%
	WEKA	98.00%	98.03%	98.06%	98.10%	98.07%	98.22%	94.84%	91.25%	89.32%	89.37%	89.93%	90.69%		89.32%	98.22%	8.91%
Прецизност	MATLAB	83.47%	94.18%	89.19%	78.70%	89.51%	95.03%	84.16%	79.59%	78.23%	78.02%	77.91%	77.86%	**0.78**	77.86%	95.03%	17.17%
	WEKA	98.20%	98.50%	98.10%	97.80%	98.10%	98.30%	95.20%	91.40%	90.50%	90.40%	90.50%	90.90%		90.40%	98.50%	8.10%
Чувствителност	MATLAB	96.00%	98.58%	97.38%	94.84%	97.46%	98.89%	96.17%	93.92%	94.17%	93.99%	93.97%	94.15%	**0.91**	93.92%	98.89%	4.97%
	WEKA	97.10%	98.10%	95.70%	94.50%	95.60%	98.20%	94.80%	91.20%	89.30%	89.40%	89.90%	90.70%		89.30%	98.20%	8.90%
Ф мярка	MATLAB	89.30%	96.33%	93.11%	86.02%	93.31%	96.92%	89.76%	86.16%	85.46%	85.26%	85.19%	85.23%	**0.85**	85.19%	96.92%	11.73%
	WEKA	96.60%	97.60%	97.10%	94.20%	97.10%	96.00%	94.70%	91.60%	88.70%	88.80%	89.50%	90.40%		88.70%	97.60%	8.90%
Класификатор		SVM	SVM	SVM	SVM	SVM	SVM	SVM	Bagged	Boosted	Boosted	Boosted	Boosted				



Фиг. 4.19 Графики на метрики получени за възраст 66-99 години–III група

4.8.2. Обобщение, анализ и изводи от проведените изследвания

От направените изследвания по възраст и получени експериментални резултати (таблица 4.8, таблица 4.9, таблица 4.10, фиг.4.17, фиг. 4.18 и фиг. 4.19) са направени следните изводи:

I група

- Изчисленият коефициент на корелация за точност между резултатите, получени от MATLAB и WEKA е със стойност от **0.5249** т.е. установена е значителна зависимост между двете изследвани поредици;
- Резултатите от изчислената метриката за **Точност** показват, че диапазона между минимална и максимална стойност за MATLAB е малък, докато при WEKA той е значителен;
- От диаграмата фиг.4.17 за **Точност**, определени характеристични описатели в MATLAB, в началото е по-висока след което точността при WEKA е доминираща.
- Получената стойност за коефициента на корелация за Прецизност е **0.4503,** т.е. тя може да се определи като умерена зависимост;
- От Фиг. 4.17 се наблюдава, че **Прецизността** с WEKA е по-висока сравнение с тази на MATLAB;
- Коефициентът на корелация за **Чувствителност** е **0.4492**, което определя умерена зависимост;
- За метриката **Чувствителност,** от получените резултати се констатира, че в WEKA те са по-високи, в сравнение с MATLAB;
- Изчисленият коефициент на корелация за **F-мярка** е **0.4343** т.е. установява се умерена зависимост;

- За **F-мярка** от визуализираните резултати се забелязва, че разликите между минимална и максимална стойност са големи.

II група

- За метриката **Точност** от получените резултати се установява, че разликите между минимална и максимална стойност е по-голяма за WEKA в сравнение с MATLAB;
- Коефициентът на корелация от таблица 4.9, изчислен от стойностите, получени от MATLAB и WEKA за метриката **Точност** е **0.8582** зависимостта е определена като значителна;
- От диаграмата за **Точност** се забелязва, че резултатите за MATLAB и WEKA до 9-ти описател са с близки стойности, а от описател 10 MATLAB е с по-високи резултати;
- От направеният анализ на получените резултати по отношение на разликата между минимална и максимална стойност за **Прецизност** в MATLAB е получен диапазон с голяма разлика, сравнение с WEKA;
- Стойността за коефициента на корелация за показател **Прецизност** е **0.739,** т.е. зависимостта е значителна;
- От диаграмата на фиг. 4.18 се установява, че **Прецизността** определена с WEKA е по-висока сравнение с тази на MATLAB;
- Относно показател **Чувствителност**, от стойностите на получените резултати се забелязва, че WEKA е с малко по-високи стойности, в сравнение с MATLAB;
- Изчисленият коефициент на корелация, за **Чувствителност** е **0.8559,** т.е. получената зависимост е силна;
- За **F-мярка** от данните на резултатите се установява, че разликата между минимална и максимална стойност е голяма за MATLAB, а за WEKA е малка;
- Коефициентът на корелация за показател **F-мярка** е **0.6653** т.е. се установява значителна зависимост.

III група

- За метрика **Точност** диапазоните между минималната и максимална стойност са малки и до известна степен се препокриват;
- Изчисленият коефициент на корелация за точността между резултатите получени от MATLAB и WEKA, е с висока стойност **0.8232.** Получената зависимост е силна;
- От направеният сравнителен анализ на резултатите за дванадесетте изследвани описатели се установява, че разликата между минималната и максимална стойност за **Прецизност** е незначително увеличена в сравнение с тази на точността;
- Корелационият коефицинет е със стойност **0.7846**, зависимостта е силна;
- За показател **Чувствителност**, резултатите получени с MATLAB определено са по-високи и диапазона на изменение е **4.97%** в сравнение с WEKA;
- Коефициентът на корелация е висок **0.9078**, т.е. получената зависимост е силна;
- За показател **F-мярка** изменението на резултатите е динамично и съответно коефициента на корелация между тях е **0.8451**.

4.9. Изводи от валидацията на характеристичните описатели

В настоящата глава от дисертационния труд е реализиран третият етап от технологичния подход за редуцуране и изследване на характеристични описатели за високи нива на дистрес. О получените резултати са могат да се направят следните изводи:

- Сравнителният анализ на резултатите от изследванията на предоставените клинични данни посредством MATLAB и WEKA показва силна корелационна зависимост, което може да се приеме като взаимно валидиране и съответно висока степен на достоверност при използването им в клиничната практика;
- Диапазонът на изменение според изчислените стойности, определен от разликата между макимална и минимална стойност е незначителен , което до известна степен може да се приеме като диапазон на очаквани резултати от клинични изследвания;
- Проведените допълнителни изследвания на пациентите по пол с MATLAB и WEKA отново взаимно се потвърждават, като получените метрики за точност, прецизност, чувствителност и F-мярка са високи. Това се приема за основание, че изследваните характеристични описатели могат да бъдат използвани в клиничната практика в качеството на инструментално средство за диагностика на високи нива на дистрес;
- Резултатите от проведените изследвания на характеристичните описатели, приложени върху три основни групи пациенти показват достатъчно високи корелационна зависимост. Диапазонът на тяхното изменение /максимална и минимална граница/ е значително по-голям в сравнение с останалите изследвания, което съответства на възрастовите изменения и различната им склонност за преминаване в дистресово състояние след поставена диагноза на онкологично заболяване;

Резултатите от проведените изследвания посредством MATLAB и WEKA взаимно се потвърждават на базата на високата корелационна зависимост, независимо от ясно изразения субективен характер на клиничните данни. Изводът е в сила и за изследвания при сегментиране на пациентите по пол и възраст. Диапазонът на изменение на изчислените метрики се оценява като незначителен, независимо, че при пациенти от третата група той е увеличен.

ЗАКЛЮЧЕНИЕ

От получените експериментални резултати и изводи могат да се направят следните заключения:

Необходими са широкомащабни изследвания на признаците за диагностициране и оценка нивото на дистрес, посредством съвременни теоретични достижения и клинични резултати, базирани на средствата за машинно обучение. Насочеността на основният проблем и целта на дисертационния труд – синтез и изследване полезността на характеристични описатели за нивото на дистрес, с цел създаване на инструментални средства в помощ на клиничните практики за диагностициране на дистрес.

В дисертационния труд е разработен технологичен подход, при които поетапно се синтезират и изследват характеристични описатели на дистрес, базирани на множество от комбинации на признаци, основаващи се на субективната самооценка на изследваните пациенти и тяхното редуциране до приемливи за клиничната практика брой на характеристични описатели. В разработения технологичен подход са използвани математически и програмни средства съобразени с диапазона и количеството на изследваните признаци, поетапно както следва:

Предварителната обработка включва изчисления с общ характер върху предоставените клинични данни, посредством EXCEL от MS Office с цел получаването на начални данни за първично редуциране на описателите и за следващите изследвания.

Провеждане на машинно обучение, посредством MATLAB и вторично редуциране на характеристичните описатели на чрез сравнителна оценка на базовите метрики за точност, прецизност, чувствителност и F-метрика.

Като се има в предвид важността на получените описатели и тяхното клинично приложение е направена вторична валидация, посредством програмната среда WEKA на основните метрики на разработените характеристични описатели за нивото на дистрес, което допълнително потвърждава достоверността на получените резултати. На базата на разработените и потвърдени характеристични описатели е проведено изследване на предоставените клинични данни, посредством MATLAB и WEKA, катп е направено сегментиране по пол и възраст.

В заключение, отчитайки високите стойности, на основни метрики при детекция на дистрес, вторичната валидация и допълнително изчислените корелационни зависимости между резултатите от паралелните изследвания, проведени с различни програмни средства, се изказва обосновано предположение, че предложените характеристични описатели могат да бъдат използвани в клиничната практика, като допълнително инструментално средство при диагностика на дистрес. За проверка на това предположение е необходимо да се сформира интердисциплинарен екип, който да проведе изследвания с реални пациенти, което излиза извън обхвата на настоящия дисертационен труд.

ПРИНОСИ ПО ДИСЕРТАЦИЯТА

НАУЧНО – ПРИЛОЖНИ ПРИНОСИ

1. Предложени са пет алгоритъма за синтез на характеристични описатели. Чрез тях са синтезирани голям брой характеристични описатели за откриване на високи нива на дистрес.
2. Получена е оценка на приложимостта на голям брой характеристични описатели при разпознаване на високи нива на дистрес според критерий. Оценката е използвана за редуциране на първоначалния набор от 32 описатели до 12.
3. Получена е оценка на полезността за голям брой характеристични описатели, използвани в комбинация с пет разнородни метода за класификация. Въз основа на получените резултати са предложени описатели, които имат потенциал да служат като допълнителни индикатори в клиничните практики за откриване на високи нива на дистрес.

ПРИЛОЖНИ ПРИНОСИ

1. Предложен е подход за извличане на значими характеристични описатели от голям брой признаци, получени от „*Скрининг инструмент за самооценка на дистрес“*. За целта е използвано и подходящо програмно осигуряване.
2. Получени и валидирани са множество резултати от експериментални изследвания на характеристичните описатели в комбинация с пет категории класификатори. Предложената методика и общата технологична рамка за откриване на високи нива на дистрес създават предпоставки за разработването на специализирано програмно осигуряване в помощ на психоонколозите.

ПУБЛИКАЦИИ ПО ТЕМАТА НА ДИСЕРТАЦИОННИЯ ТРУД

1. **Гинка Маринова[1]**, Мая Тодорова[2], „Обзор на техники за извличане на данни в онкологията и психоонкологията“, списание „Компютърни науки и технологии“, ISSN 1312-3335, Година XVIII, Брой 1/2020, стр. 141-146.
2. **Гинка Маринова**, “Машинното обучение за изследване на дистрес при пациенти с онкологични заболявания“ списание „Компютърни науки и технологии“, ISSN 1312-3335, Година XVIII, Брой 1/2020, стр. 133-140.
3. Мая Тодорова[1], Нели Калчева[2], **Гинка Маринова[3]**, Недялко Николов[4], “Обзор и класификация на методи и задачи в Data Mining“ VIII International Scientific Conference “ Engineering. Technologies. Education. Safety”, 08-11.06.2020, Borovets, Bulgaria, ISSN 2535-0315, Volume 2, стр. 70-80.
4. **Ginka Marinova[1]**, Todor Ganchev[1] and Nedyalko Nikolov[2], “Synthesis of Characteristic Descriptors for the Detection of Distress” International Conference on Biomedical Innovations and Applications – BIA, 24-27 Sept., 2020, Varna, Bulgaria, DOI: 10.1109/BIA50171.2020.9244488, Scopus
5. **Ginka Marinova[1]**, Maya Todorova[2], „Classification Tasks Solving with Machine Learning Methods”, XXIX International Scientific Conference Electronics, September 16 - 18, 2020, Sozopol, Bulgaria, DOI: 10.119/ET50336.2020.9238218 Scopus
6. **Ginka Marinova[1]**, Todor Ganchev[1] and Nedyalko Nikolov[2], „Application of machine learning methods for the prediction of distress in patients with oncological diseases“ Годишник на Технически университет-Варна, 4(2), 130-137. https://doi.org/10.29114/ajtuv.vol4.iss2.204

Благодарности на:

Изказвам благодарности на доц. д-р инж. Недялко Николов, проф. д-р инж. Тодор Ганчев, доц. д-р. инж. Виолета Божикова, на колегите от катедра „Софтуерни и интернет технологии" и от катедра „Компютърни науки и технологии" ТУ-Варна, за оказаната помощ и подкрепа при разработване на настоящия дисертационен труд.

ABSTRACT

Dissertation Title: Research of methods for recognition of distress from psychological tests through using of the machine learning
of the Requirement for the Degree Doctor of Philosophy
by Ginka Kaleva Marinova

The aim of the dissertation is to develop a technological approach for the synthesis of characteristic descriptors to diagnose high levels of distress, and consistently reduce the composition of descriptors and reduce them to acceptable for medical practice composition.

Chapter 1 describes problems related to the causes of distress, outlines basic terms of medical diagnosis, and psycho-oncology. Technical tools for registration, storage, and processing of medical data are addressed. An analysis of the mathematical and experimental methods and algorithms for processing the results for psychological distress is made and critical analysis of previous developments on the topic is made additionally. A classification of the groups of methods is proposed and a comparative analysis of their advantages and disadvantages is presented.

In Chapter 2 the first stage of the technological approach for synthesis, research, and argumentative reduction of the characteristic descriptors for the diagnosis of high levels of distress is addressed. The distribution of the features is presented as an initial assessment of their impact on the level of distress. The correlation coefficient between the features themselves is calculated to establish the existence of a link between them, which is based on the same indicators. A comparative analysis of different methods and algorithms for classification is made in order to assess their applicability in the detection of distress based on data from standard questionnaires for the assessment of distress.

Chapter 3 presents the second stage of the developed technological approach for synthesis and research of characteristic descriptors of high levels of distress. The main task is to synthesize characteristic descriptors based on the information extracted from the previous chapter. The descriptors obtained by means of machine learning methods are studied.

Chapter 4 consists of the validation of the developed descriptors from the previous chapter, reduction of the number of descriptors, and their examination for the cases when the patients are grouped by sex and age, is made. The evaluation of the obtained indicators based on machine learning is made.