Primena tehnika mašinskog učenja u klasifikaciji pacijenata posle infarkta miokarda

Slađana Jovanović
Preduzeće za telekomunikacije
"Telekom Srbija", A.D.
Beograd, Srbija
sladjanajo@telekom.rs

Dragana Bajić
Fakultet tehničkih nauka
Univerzitet u Novom Sadu
Novi Sad, Srbija
dragana.bajic@gmail.com

Milan Jovanović
"Endava".
Beograd, Srbija

milan.jovanovic@endava.com

Branislav Milovanović
Medicinski fakultet
Univerzitet u Beogradu
Beograd, Srbija
branislav_milovanovic@vektor.net

Sažetak — **U radu su analizirani kardiovaskularni parametri snimljeni u grupi pacijenata sa dijagnozom infarkta miokarda i kontrolnoj grupi zdravih pacijenata. Cilj istraživanja je bio da se načini model koji će rešavati problem binarne klasifikacije, predikcije rizika infarkta miokarda i uticaja pojedinačnih parametara na odluku o pripadnosti klasi. U radu je dat pregled korišćenih modela koji su bazirani na 4 različite tehnike mašinskog učenja: *Decision Tree, Random Forest, Support Vector Machine* i *Artificial Neural Networks*. Komparativnom analizom rezultata dobijena je najveća ukupna tačnost predikcije od 95,58% za model baziran na *Random Forest* tehnici mašinskog učenja.**

Ključne riječi – infarkt miokarda; biosignali; tehnike mašinskog učenja; klasifikacija; predikcija; procena rizika

I. UVOD

Svetska zdravstvena organizacija svrstala je Srbiju na treće mesto u svetu po broju umrlih od bolesti srca i krvnih sudova [1]. Tokom 2017. godine od ovih bolesti u Srbiji je preminulo 53.668 ljudi. Stoga se u Srbiji, kao i u čitavom svetu, sprovodi veliki broj istraživanja u oblasti kardiologije. Korišćenje veštačke inteligencije i tehnika mašinskog učenja treba da obezbedi set alata koji će unaprediti rad kardiologa. Osim toga, u vreme intenzivnog razvoja *big data* i *data rich* tehnologija, poput povezivanja baza medicinskih podataka, tehnologije sekvenciranja celog genoma, biometričkih podataka registrovanih senzorima koji će se prenositi mobilnim mrežama, postavljaju se zahtevi da se informacije iz različitih oblasti biomedicine, pa i kardiologije, analiziraju na jedan nov analitički način, kroz scenarije koji se ne mogu realizovati tradicionalnim statističkim metodama.

Radovi iz oblasti primene tehnika mašinskog učenja u kardiologiji počinju da se objavljuju od 1995. godine [2, 3], ali je oblast postala naročito popularna poslednjih godina. Analize su bazirane na različitim ulaznim podacima. Tako su neki radovi koji su imali za cilj procenu rizika bazirani na podacima iz medicinskih kartona [4] poput godina, visine, težine, pritiska, podataka o bolestima, pušenju i drugim medicinski relevantnim podacima, dok su neki radovi uzimali u obzir podatke snimane od pacijenata poput podataka iz elektrokardiograma [5,6] ili druge kliničke podatke poput podataka iz angiograma i koronarnog CT-a [7]. Kako je u kliničkim ispitivanjima veliki problem veličina baze, radovi su se bavili komparativnim analizima tehnika koje daju najbolje rezultate sa malim brojem podataka [8] i poređenjem sa klasičnim statističkim metodama koja su pokazala prednost tehnika mašinskog učenja [9]. Radovi [10-12] daju pregled tehnika mašinskog učenja koje su korišćene u kardiologiji, uz poređenje prednosti i nedostataka tehnika, ali uz jedinstven zaključak, da veštačku inteligenciju i tehnike mašinkog učenja kao granu ove nauke, treba inkorporirati u kliničku praksu. To će omogućiti tumačenje i dublje razumevanje velikog broja medicinskih podataka uz otkrivanje veza među njima koje druge tehnike ne mogu da otkriju.

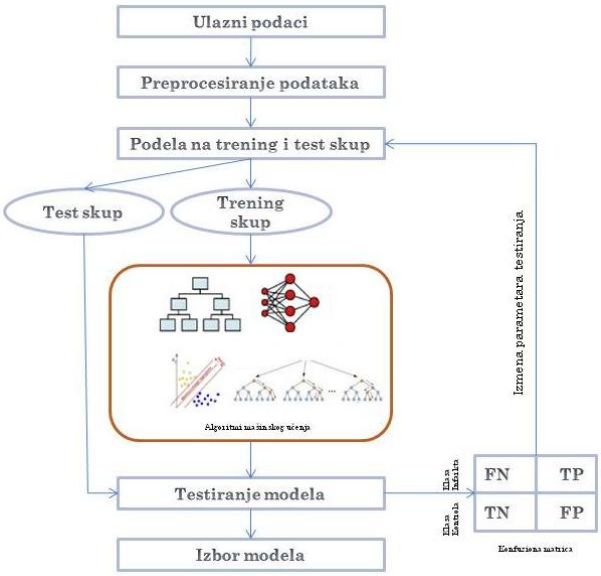
Ovo istraživanje se bazira na podacima dobijenim u standardnim dijagnostičkim procedurama u kliničkim ispitivanjima, snimanjem električnog potencijala srca, elektrokardiograma (EKG), i podataka iz 24h holter monitoringa EKG-a i pritiska. Značajno je što su ovim podacima dodati i podaci sa *Task Force Monitora* (TFM), aparata koji snima hemodinamičke parametre. Osim toga ovaj aparat proračunava i barorefleksnu osetljivost, parametar koji je izuzetno važan u proceni funkcionalnosti kardiovaskularnog i autonomnog nervnog sistema. U radu je primenjena komparativna analiza različitih tehnika mašinskog učenja. Iako je primarni cilj bio rešavanje problema binarne klasifikacije između kontrolne grupe zdravih osoba i obolelih od infarkta miokarda, razvijeni model je omogućio i procenu rizika za nove pacijente i analizu uticaja parametara na odluku o klasifikaciji.

Istraživanja su finansirana grantovima TR32040 i III41013 Ministarstva prosvete, nauke i tehnološkog razvoja Republike Srbije i pod podkroviteljstvom su EU COST - Action CA15104 "Inclusive Radio Communication Networks for 5G and beyond" - SEWG-IoT: Internet -of - Things for Health.

II. METODOLOGIJA I CILJEVI ISTRAŽIVANJA

A. *Metodologija istraživanja*

Istraživanje je sprovedeno u skladu sa metodologijom CRISP-DM (*Cross Industry Standard Process for Data Mining*) [13]. Ova metodologija propisuje 6 koraka: 1) razumevanje problema i definisanje cilja istraživanja; 2) razumevanje podataka; 3) preprocesiranje podataka; 4) razvoj modela korišćenjem komparabilnih analitičkih tehnika; 5) evaluacija rešenja i komparativna procena validnosti modela; 6) primena modela u procesima odlučivanja. Sprovođenje svakog od ovih koraka obezbeđuje da se na sistematičan način sprovedu istraživanja što doprinosi tačnosti dobijenih rezultata. Prva dva koraka su od izuzetne važnosti za shvatanje problema. Ovo je posebno izraženo u multidisciplinarnim istraživanjima, kada tim stručnjaka iz različitih oblasti usaglašava svoje delovanje kako bi postigli definisani cilj istraživanja. Preprocesiranje podataka predstavlja vremenski možda najzahtevniju fazu u okviru metodologije. Ove faze predstavljaju preduslov da se kroz naredne faze izrade modela, komparativne analize i primene modela, dobiju validni rezultati istraživanja. Šematski prikaz procedure istraživanja, usklađene sa metodologijom CRISP-DM prikazan je na slici 1.



Slika 1. Procedura istraživanja

.

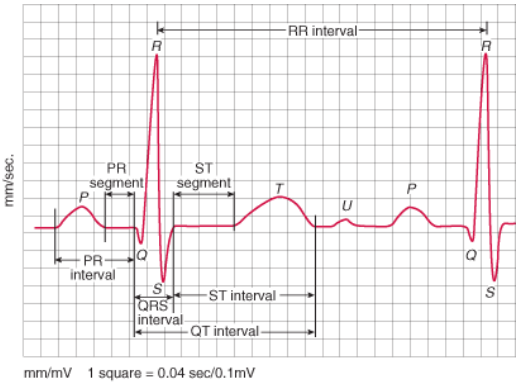
B. Cilj istraživanja

Cilj istraživanja je bio da se sačini model složenih zavisnosti između izabranih kardiovaskularnih parametara u grupi pacijenata koji su preležali infarkt miokarda i kontrolnoj grupi zdravih osoba. Model bi se koristio u svrhu klasifikacije, predikcije verovatnoće pripadnosti jednoj od klasa za novotestiranu osobu i određivanje uticaja kardiovaskularnih parametara na odluku o pripadnosti klasi.

III. Kardiovaskularni parametri kao ulazni podaci

Kardiovaskularni parametri pacijenata i zdravih osoba su snimljeni u neurokardiološkoj laboratoriji Univerzitetskog kliničkog centra "Bežanijska kosa" u Beogradu, u Srbiji. Pacijenti su bili primljeni na koronarno odeljenje nakon preležanog infarkta miokarda.

Broj pacijenata sa preležanim infarktom miokarda koji je učestvovao u istraživanjima je 499, dok je broj zdravih osoba čiji su parametri snimljeni 102. Ukupan broj parametara je bio 49. Parametri su snimani na različitim uređajima i jedan od ciljeva rada je bio da se uporedi njihov uticaj u procesu klasifikacije. 12 parametara je snimano ili proračunato korišćenjem elektrokardiografa, uređaja koji beleži električnu aktivnost srca u vremenu: QTc, QT, PR, QRS, P, RR, Paxis, QRSaxis, Taxis, SDRR, PNN50, RMSSD. 12 parametara je snimano ili proračunato iz 24h holtera EKG-a: Average HR, SDNN, SDANN Index, SDNN Index, RMSSD, PNN50, Power, VLF, LF, HF, ULF, LFHF. 3 parametra su snimana ili proračunata iz 24h holtera pritiska: SBP, DBP, pulsni pritisak. 22 parametra su snimana ili proračunata aparatom TFM: HR, SBP, DBP, MBP, LFnuRR, HFnuRR, VLF_RRI, LF_RRI, HF_RRI, PSD_RRI, LFHF_RRI, LFHF, LFnu_RRI_DBP, HFnu_RRI_DBP, VLF_DBP, LF_DBP, HF_DBP, PSD_DBP, LFHF_DBP, LFHF_PBV_DBP, BRS, BEI.



Slika 2. Elektrokardiogram (slika preuzeta sa http://lifeinthefastlane.com/ecg-st-segment-evaluation/)

Na slici 2 je prikazan elektrokardiogram na kome su označeni talas P, QRS kompleks i talas T, kao i karakteristični intervali čije se trajanje meri. U analizi su korišćeni i parametri koji ukazuju na pravac električnog vektora depolarizacije pretkomore, komore i repolarizacije komore, respektivno Paxes, QRSaxes i Taxes. U dijagnostici kardiovaskularnih oboljenja koriste se parametri koji ukazuju na varijabilnost u vremenskim intervalima između dva srčana otkucaja, i nazivaju se parametrima varijabilnosti srčanog ritma. Imaju veliki dijagnostički i prognostički značaj i kroz fiziološko tumačenje ukazuju na aktivnost autonomnog nervnog sistema. Proračunavaju se u vremenskom ili frekvencijskom domenu. Među najznačajnijim parametrima koji se proračunavaju u vremenskom domenu su SDNN (standardna devijacija NN intervala), RMSSD (srednja kvadratna vrednosti uzastopnih razlika), SDSD (standardna devijacija uzastopnih razlika) i PNN50 (procenat susednih

parova NN intervala koji se razlikuju za više od 50ms). Metode koje se koriste u frekvencijskom domenu proračunavaju broj NN intervala koji pripadaju odgovarajućim frekvencijskim opsezima. Za ljudski organizam opseg visokih frekvencija (HF) je od 0,15 do 0,4Hz, niskih frekvencija (LF) je od 0,04 do 0,15 Hz i opseg vrlo niskih frekvencija je (VLF) je od 0,0033 do 0,04 Hz. Parametri u frekvencijskom domenu uključuju i odnose snaga u opsezima niskih i visokih frekvencija. TFM je uređaj specifičan po tome što proračunava barorefleksnu osetljivost (BRS) i indeks barorefleksne efikasnosti (BEI), parametre koji direktno ukazuju na funkcionisanje barorefleksnog mehanizma, jednog od najznačajnijih fizioloških mehanizama za regulaciju srčanog pritiska kroz modulaciju srčanog ritma delovanjem autonomnog nervnog sistema.

Prvog dana nakon infarkta miokarda mereni su kardiovaskularni parametri korišćenjem EKG-a. Parametri koji su dobijeni korišćenjem 24h Holter monitoringom EKG-a i pritiska i TFM-om snimani su dve nedelje nakon preležanog infarkta miokarda. Svi eksperimentalni protokoli su odobreni od strane Naučnog etičkog komiteta Univerzitetskog kliničkog centra "Bežanijska kosa", licenca broj 1039/3. Svi učesnici su bili potpuno informisani o istraživanju i dali su pisanu saglasnost u skladu sa Helsinškom deklaracijom.

U cilju razumevanja podataka i pripreme za dalje analize rađena je njihova statistička obrada u smislu proračuna opsega, srednjih vrednosti i standardnih devijacija za svaki od parametara. Grafička predstava putem kutijastog dijagrama, *Box Plot*-a korišćena je za identifikaciju ekstremnih vrednosti, *outlier*-a. *Pearsson*-ovom linearnom i *Spearman*-ovom rank korelacijom vršena je analiza linerane, odnosno, nelinearne zavisnosti izabranih parametara. Uočena je nebalansiranost u broju instanci dve klase, jer klasa MI (osoba sa preležanim infarktom miokarda) sadrži 4,89 puta više instanci od kontrolne klase (zdrave osobe).

Prilikom preprocesiranja podataka vrlo strogo je postavljen uslov da se ne naruše zakonitosti i međuveze između kardiovaskularnih parametara jedne osobe. Stoga su odbačeni svi rezultati osoba koje su imale više od 3 nedostajućih vrednosti. Broj korigovanih *outlier*-a u izabranim instancama je bio maksimalno 3 po parametru, i to samo za *outlier*-e koji su imali značajna, najčešće višestruka odstupanja od srednjih vrednosti raspodele konkretnog parametra. U slučaju dopune nedostajućih vrednosti i korekcije vrednost *outlier*-a korišćene su srednje vrednosti parametra u pripadajućoj klasi. U slučaju da tehnika analize to zahteva rađene su transformacije podataka u isti numerički opseg uz očuvanje veza između parametara. Zbog očuvanja prirode multivarijantne zavisnosti između parametara nije rađeno balansiranje broja instanci u klasama dodavanjem, odnosno, redukcijom broja instanci. Problem nebalansiranosti veličine klasa rešen je kroz razvoj modela u okviru implementacije tehnika mašinskog učenja.

IV. RAZVOJ MODELA

A. *Pregled korišćenih tehnika mašinskog učenja*

Rezultati sprovedenih analiza i preprocesiranja podataka korišćeni su prilikom izbora tehnika mašinskog učenja za generisanje modela multivarijantne raspodele kardiovaskularnih parametara. Prilikom izbora tehnika treba imati u vidu da one mogu biti izuzetno osetljive na postojanje nedostajućih podataka, *outlier*-a, linearne zavisnosti raspodele parametara ili broja poramatera koje karakterišu jednu instancu. Zbog uočene linearne zavisnosti nekih od parametara nije korišćena tehnika linearne regresije. Preprocesiranjem podataka izbegnuti su problemi nedostajućih podataka, različitih opsega podataka (za tehnike koje to zahtevaju), i umanjen je problem *outlier*-a zamenom njihovih najekstremnijih vrednosti. Problem nebalansiranog broja instanci u klasama delimično je rešen prilagođenjem modela baziranih na DT, RF i SVM tehnikama u smislu minimizovanja troškova pogrešne klasifikacije u fazi učenja. Time su primenjeni klasifikatori postali *Cost Sensitive* klasifikatori osetljivi na troškove [14]. U modelu baziranom na ANN tehnici nije uzeta u obzir nebalansiranost klasa.

Ovo je omogućilo da se u istraživanju koriste različite tehnike mašinskog učenja kako bi se kroz komparativnu analizu rezultata predikcije odabrao optimalan model. Odabrane četiri tehnike su među najpoznatijim i najčešće korišćenim tehnikama mašinskog učenja u medicinskim istraživanjima [10-12].

Decision Tree (DT) je tehnika mašinskog učenja koja je vrlo popularna usled jednostavne i intuitivne interpretacije [15]. Bazira se na nizu podela podataka na osnovu proračuna određene metrike kojom se ocenjuje relevantnost parametra izabranog za granjanje. U radu smo koristili dva različita algoritma: ID3 (*Iterative Dichotomizer* 3) koji koristi entropiju i informacioni dobitak kao metriku, i CART (*Classification and Regression Trees*) koji koristi Gini indeks kao metriku.

Random Forest (RF) je tehnika mašinskog učenja koja spada u klasu ansambl metoda [16]. Ansambl se sastoji od blokova – DT koji predstavljaju gradivne jedinice RF tehnike. Blokovi se prave od podskupa ukupnog broja elemenata i podskupa parametara. Kombinacijom rezultata velikog broja blokova ukupan rezultat dobija na tačnosti. Kada je u pitanju rešavanje problema klasifikacije, rezultat se dobija preglasavanjem između pojedinačnih rezultata u okviru svakog od blokova.

Support Vector Machines (SVM) pripadaju familiji generalizovanih linearnih modela baziranih na linearnoj kombinaciji parametara [17]. Tehnika je bazirana na predstavljanju podataka u vektorskom prostoru i nalaženju hiper ravni koja najbolje razdvaja podatke različitih klasa. Za rešavanje problema klasifikacije kada klase ne mogu da se razdvoje linearnom funkcijom, koristi se nelinearna kernelova funkcija kako bi se ulazni podaci transformisali u neki višedimenzionalni prostor u kome je skup podataka za trening linearno razdvojiv.

Artificial Neural Networks (ANN) su jedna od najpoznatijih i najviše korišćenih tehnika mašinskog učenja [18]. Struktura ANN oponaša strukturu nervnog sistema, mehanizam prenošenja neuralnog impulsa neuralnim putanjama i način donošenja odluke o ishodu. U ovom radu korišćena je *Multi-Layer Perceptron* (MLP) ANN sa povratnom propagacijom. Ona se sastoji od računarskih jedinica (neurona, perceptrona) raspoređenih u više slojeva i međusobno povezanih vezama sa težinskim koeficijentima. U okviru neurona je korišćena sigmoidalna aktivaciona funkcija. Proces učenja je značio podešavanje težinskih koeficijenata veze između neurona kroz mininimizaciju greške.

B. *Obučavanje i testiranje modela*

U obučavanju modela korišćene su nadgledane tehnike mašinskog učenja. U svojoj osnovi one imaju statističke teorije koje kroz analizu podataka prepoznaju obrasce i time "obučavaju" model da postigne cilj predikcije za svaku novu instancu. U radu je trening vršen na skupu od 70% podataka. Na testnim podacima koji čine 30% od ukupnog broja podataka proveravana je tačnost predikcije modela. U iterativnim procesima vršena je izmena konfiguracije modela i izmena parametara koji su specifični za svaku od tehnika mašinskog učenja. Kroz veliki broj iteracija izvršen je izbor optimalnih modela koji su dali sveukupno najbolje rezultate predikcije pripadnosti klase.

Modeli generisani u sve 4 tehnike mašinskog učenja postigli su tačnost od 100% na skupu podataka za trening. Predikcije u okviru testnog skupa podataka su korišćene kao kriterijum za izbor modela. U testnom skupu broj osoba koje pripadaju klasi obolelih od infarkta miokarda označili smo kao Pozitivna klasa (P), a broj osoba u kontrolnoj klasi zdravih osoba kao Negativna klasa (N). Broj tačno prediktovanih pripadnika klasi označili smo kao *True* Pozitivni (TP), odnosno, kao *True* Negativni (TN) za klasu P odnosno, N. Broj pogrešno prediktovanih pripadnika klasi označili smo kao *False* Pozitivni (FP), odnosno, kao *False* Negativni (FN) za klasu P odnosno, N. Senzitivnost, specifičnost, vrednost pozitivne predikcije, vrednost negativne predikcije, ukupna tačnost predikcije, *precision*, *recall* i *F1 score* na testnom skupu proračunati su na sledeći način [19]:

$$senzitivnost = {TP}/{(TP + FN)}$$

$$specifičnost = {TN}/{(TN + FP)}$$

$$pozitivna\ predikcija = {TP}/{(TP + FP)}$$

$$negativna\ predikcija = {TN}/{(TN + FN)}$$

$$tačnost\ predikcije = {(TP + TN)}/{(P + N)}$$

$$precision\ za\ klasu\ N = {TN}/{(TN + FN)}$$

$$recall\ za\ klasu\ N = {TN}/{(TN + FP)}$$

$$F1\ score = {2 * precision * recall}/{(precision + recall)}$$

U tabeli I prikazane su vrednosti ovih parametara, osim parametara *precision* i *recall*. *F1 score* je posebno značajan parametar kod nebalansiranih klasa podataka i ukazuje na balans između parametara *precision* i *recall*. On predstavlja meru osetljivosti klasifikatora i proračunat je za obe klase podataka.

Tabela I ukazuje da je model baziran na RF tehnici mašinskog učenja pokazao najbolje rezultate. Modeli bazirani na SVM i ANN tehnikama mašinskog učenja pokazali su vrlo slične performanse. Primenjeno preprocesiranje omogućilo je da se eliminišu specifične slabosti pojedinačnih tehnika u odnosu na karakteristike ulaznih podataka, tako da su do izražaja došle njihove dobre osobine. Očekivano, najlošije rezultate je imao model baziran na DT. Ova jednostavna tehnika je vrlo osetljiva na broj parametara koje karakterišu instancu. U praksi se ovaj problem rešava uvođenjem ansambl metoda, poput RF tehnike. Ono što su pokazali rezultati za F1 score u Tabeli I je da problem nebalansiranosti klasa podataka nije potpuno rešen primenjenim tehnikama.

TABELA I. KOMPARATIVAN PREGLED REZULTATA PREDIKCIJE NA TESTNOM SKUPU PODATAKA

	Senzitivnost	Specifičnost	Pozitivna predikcija	Negativna predikcija	Tačnost predikcije	F1 za klasu P	F1 za klasu N
DT	93,15	77,14	94,44	72,97	90,05	0,938	0,750
RF	97,92	86,49	96,57	91,43	95,58	0,972	0,889
SVM	97,163	82,5	95,14	89.19	93,92	0,961	0,857
ANN	96,53	86,49	96,53	86,49	94,47	0,965	0,865

V. IMPLEMENTACIJA MODELA

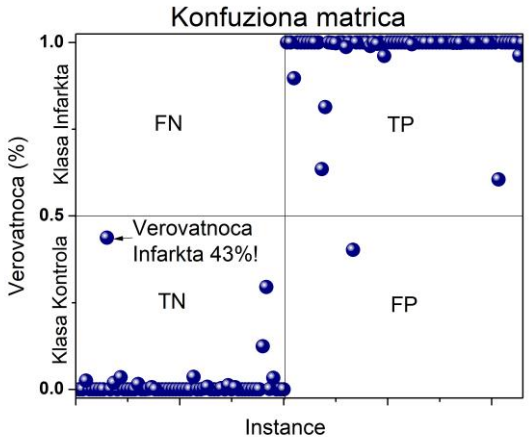
A. *Predikcija verovatnoće pripadnosti klasi*

Kako se ovako kreiran model baziran na RF tehnici mašinskog učenja može koristiti u praksi možemo videti na primeru osobe kojoj su izmereni kardiovaskularni prametri, a zatim je izvršen test predikcije pripadnosti klasi.

Rezultati testa su prikazani na slici 3, u vidu položaja nove instance u okviru konfuzione matrice. Može se uočiti da postoji verovatnoća od 43 procenta da će se instanca (osoba čiji su kardiovaskularni parametri mereni) naći u klasi obolelih od infarkta. Ova verovatnoća znači upozorenje da kardiovaskularni parametri ukazuju na mogućnost bolesti i potrebu preventivnog delovanja.

B. Uticaj parametara na odluku o pripadnosti klasi

Uticaj parametara na odluku o predikciji je jako značajan



Slika 3. Predikcija verovatnoće oboljevanja od Infarkta miokarda za novu instancu

segment istraživanja iz više razloga. U ovom složenom istraživanju koje kombinuje parametre dobijene merenjem različitim instrumentima, u različitim vremenskim intervalima, bilo je potrebno sagledati koje grupe parametara i koji parametri najviše utiču na odluku o pripadnosti klasi. Osim toga, ovo istraživanje može se koristiti u budućnosti, kako bi se broj parametara u analizi smanjio. Time se smanjuje negativan uticaj redundantnih parametara koji ne utiču na krajnju odluku o klasifikaciji i skraćuje procesorsko vreme obrade podataka.

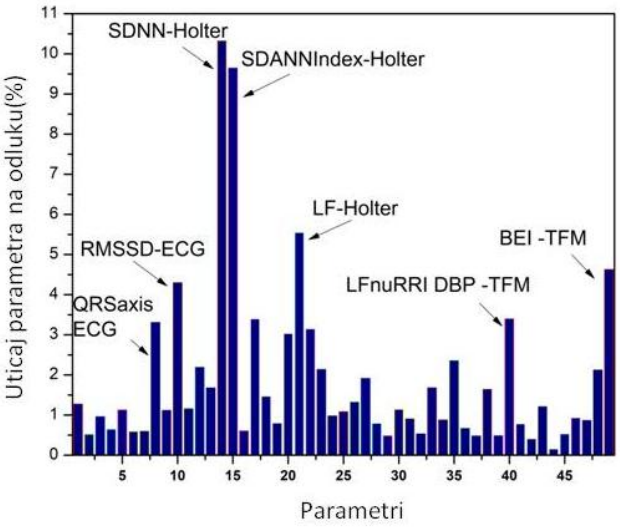
Model koji je dao najbolje rezultate je baziran na RF tehnici mašinskog učenja, te je za dati model i rađena procena uticaja parametara korišćenjem proračuna *Mean Decrease Impurity* (MDI) indeksa. MDI indeks se koristi za procenu redukcije "nečistoća" (pogrešno opredeljenih instanci za pripadnost klasi) kada se podela vrši na osnovu datog parametra u čvoru pojedinačnih DT. Formula za proračun uticaja parametra x_j MDI metodom, bazirana na Gini indeksu [20] je:

$$UP(x_j) = \frac{1}{n_{tree}}\left[1 - \sum_{k=1}^{n_{tree}} Gini(j)^k\right]$$

Vrednost Gini indeksa se proračunava za svaki od n_{tree} DT i za svaki parametar u okviru RF. Uticaj parametara prikazan je na slici 4.

VI. ZAKLJUČAK

U radu je prikazana primena četiri tehnike mašinskog učenja u generisanju modela za klasifikaciju kontrolne klase zdravih osoba i klase osoba obolelih od infarkta miokarda. Klasifikacija je vršena na osnovu kardiovaskularnih parametara snimljenih uređajima EKG, Holter EKG-a, Holter pritiska i TFM. Kroz analizu i preprocesiranje podataka otklonjeni su nedostaci iz skupova podataka (nedostajuće



Slika 4. Uticaj parametara na predikciju klase

vrednosti, *outlier*-i, različiti opsezi vrednosti podataka), što je omogućilo korišćenje tehnika mašinskog učenja koje su osetljive na ove pojave. Analizom međusobne zavisnosti parametara uočena je linearna zavisnost grupe parametara što je uticalo na izbor tehnika mašinskog učenja. Uticaj nebalansiranog broja instanci po klasama je rešen kod modela baziranih na DT, RF i SVM tehnikama njihovom modifikacijom u *Cost Sensitive* klasifikatore. Ova modifikacija nije urađena na modelu baziranom na ANN tehnici. Modeli bazirani na svakoj od tehnika su generisani kroz više iteracija u kojima su setovane različite konfiguracije i parametri koji ih karakterišu, tako da su u daljoj analizi korišćeni modeli najboljih performansi za svaku od tehnika.

Komparativnom analizom odabran je kao najbolji model baziran na RF tehnici mašinskog učenja. Generisani modeli bazirani na ANN i SVM tehnikama su sličnih performansi. Model baziran na DT tehnici je imao najlošije performanse. DT tehnika mašinskog učenja je najjednostavnija tehnika i vrlo je osetljiva na broj parametara koji karakterišu instancu, tako da su ovakvi rezultati modela očekivani. Parametar *F1 score* je ukazao da model nije u potpunosti rešio problem nebalansiranosti klasa podataka. Interesanstno je da je model baziran na ANN tehnici imao slične performanse sa modelima baziranim na *Cost Sensitive* RF i SVM tehnikama.

Prema prikazanim rezultatima može se zaključiti da model može da se primeni za klasifikaciju novih pacijenata. Poseban značaj u smislu medicinskih istraživanja je proračun uticaja parametara na odluku o pripadnosti klasi. Oni su ukazali na veliki uticaj parametara snimljenih 24h holter monitoringom, SDNN i SDANN Index. Ovi parametri su dijagnostički vrlo značajni. Dobijaju se analizom varijabilnosti srčanog ritma u vremenskom domenu. Time su potvrđena teorijska očekivanja. Izdvojili su se i parametri niskofrekventnog spektra, kao i barorefleksni indeks efikasnosti, BEI. Može se reći da su najveći uticaj na odluku imali parametri sa 24h holter monitoringa, zatim sa TFM-a, i tek na kraju sa EKG-a.

Naredna istraživanja ići će u pravcu daljeg poboljšanja postojećih modela, primene novih tehnika mašinskog učenja,

daljeg rešavanja problema nebalansiranih klasa podataka, smanjenja broja ulaznih parametara u skladu sa uticajem parametara u procesu odlučivanja, sačinjavanju posebnih modela za svaku od grupa podataka (EKG, Holter, TFM) i analizi senzitivnosti modela na promene vrednosti pojedinačnih parametara.

ZAHVALNICA

Prezentaciju ovog rada podržao je projekat TR32040 Ministarstva obrazovanja, nauke i tehnološkog republike Srbije. Rad je pod pokroviteljsktvom EU COST Akcije CA15104 "Inclusive Radio Communication Networks for 5G and beyond"- SEWG-IoT: Internet-of-Things for Health.

LITERATURA

[1] https://www.who.int/cardiovascular_diseases/en/

[2] J. Ortiz, C.G. Ghefter, C.E. Silva, R.M. Sabbatini, "One-year mortality prognosis in heart failure: a neural network approach based on echocardiographic data", J Am Coll Cardiol., 26:1586–93, 1995.

[3] F. Atienza, N. Martinez-Alzamora, J.A. De Velasco, S. Dreiseitl, L. Ohno-Machado, "Risk stratification in heart failure using artificial neural Networks", Proc AMIA Symp 2000:32–6.

[4] M.J. Kolek, A.J. Graves, M. Xu i drugi, "Evaluation of a prediction model for the development of atrial fibrillation in a repository of electronic medical records", JAMA Cardiol, vol.1, pp. 1007-1013, 2016.

[5] Q.A. Rahman, L.G. Tereshchenko, M. Kongkatong, T. Abraham, M.R. Abraham, H. Shatkay, "Utilizing ECG-based heartbeat classification for hypertrophic cardiomyopathy identification", IEEE Trans. Nanobioscience, 14(5):505-12, 2015.

[6] N. Kannathal, U.R. Acharya, C.M. Lim, P. Sadasivan, S. Krishnan, "Classification of cardiac patient states using artificial neural networks", Exp. Clin. Cardiol. Vol. 8, pp. 206–211, 2003.

[7] M. Motwani, D. Dey, D.S. Berman, "Machine learning for prediction of all-cause mortality in patients with suspected coronary artery disease: a 5-year multicentre prospective registry analysis", Eur Heart J, 38 (2017), pp. 500-507, 2017.

[8] M. Pavlou, G. Ambler, S.R. Seaman, O. Guttmann, P. Elliott, M. King, R.Z. Omar, "How to develop a more accurate risk prediction model when there are few events", BMJ, vol.351, h3868, 2015.

[9] S.F. Weng, J. Reps, J. Kai, J.M. Garibaldi, N. Qureshi, "Can machine-learning improve cardiovascular risk prediction using routine clinical data?" PLoS One, vol.12, art. no. 0174944, 2017.

[10] C. Krittanawong, H. Zhang, Z. Wang, M. Aydar, T., Kitai, "Artificial Intelligence in Precision Cardiovascular Medicine", Jour. of the Amer. Coll. of Card., Vol, 69, pp. 2657-2664, 2017.

[11] K.W. Johnson, J.T. Soto, B.S. Glicksberg, K.Shameer, R. Miotto, M. Ali, E. Ashley, J.T. Dudley, "Artificial Intelligence in Cardiology", Journal of the American College of Cardiology, vol.71, pp. 2668-2679, 2018.

[12] K. Shameer, K.W. Johnson, B.S. Glicksberg, J.T. Dudley, P.P. Sengupta, "Machine learning in cardiovascular medicine: are we there yet?", Heart, vol.114, pp. 1156-1164, 2018.

[13] C. Shearer, "The CRISP-DM model: the new blueprint for data mining", Journal of Data Warehousing Vol. 5, pp.13–22, 2000.

[14] C. Elkan, "The foundations of cost-sensitive learning", In Proceedings of the 17th international joint conference of artificial intelligence, pp. 973–978, Seattle, 2001.

[15] L. Breiman, J.H. Friedman, R.A. Olshen, C.J. Stone, "Classification and Regression Trees", Wadsworth and Brooks/Cole Advanced Books and Software, Monterey, CA, 1984.

[16] L. Breiman, "Random Forests". Machine Learning, vol.45, pp.5–32, 2001.

[17] C. Cortes, V. Vapnik, "Support vector networks", Machine Learning, vol.20, pp.273–297, 1995

[18] S. Haykin, "Neural Networks, and Learning Machines", 3rd ed.; Prentice Hall Publishing, Englewood Cliffs, pp. 1-936, New Jersey, 2008;

[19] T. Fawcett, "An Introduction to ROC Analysis", Pattern Recognition Letters. 27 (8), pp. 861–874, 2006

[20] C. Strobl, A.L.Boulesteix, T. Kneib, T. Augustin, A. Zeileis, "Conditional variable importance for random forests," *BMC Bioinformatics*, vol. 9, no. 1, article 307, 2008.

ABSTRACT

In this paper we analyzed cardiovascular parameters recorded from patients with diagnosis of myocardial infarction and healthy individuals in a control group. The aim was to develop model for solving a problem of binary classification, myocardial infarction risk prediction and parameter's influence on decision of class affiliation. The paper presents an overview of used models based on 4 different machine learning techniques: Decision tree, Random forest, Support vector machine and Artificial neural network. The comparative analysis of the obtained results resulted in the highest total precision of the prediction of 95.58% for the model based on the Random forest technique of machine learning.

MACHINE LEARNING TEHNIQUES IMPLEMENTATION FOR CLASSIFICATION OF PATIENTS AFTER MYOCARDIAL INFARCTION

Slađana Jovanović, Milan Jovanović, Dragana Bajić, Branislav Milovanović