PREPOZNAVANJE BOLEZNI NA PODLAGI VPRAŠALNIKA IN MERITEV S SENZORJI VITALNIH ZNAKOV

Maja Somrak [1,2], Anton Gradišek [1], Mitja Luštrek [1, 2], Matjaž Gams [1, 2]
1 Institut Jožef Stefan, Jamova cesta 39, 1000 Ljubljana, Slovenija
2 Mednarodna podiplomska šola Jožefa Stefana, Jamova cesta 39, 1000 Ljubljana, Slovenija
{maja.somrak, anton.gradisek, mitja.lustrek, matjaz.gams}@ijs.si

POVZETEK
V prispevku predstavljamo metodo, ki je bila razvita za prepoznavanje vrste bolezni na podlagi podatkov senzorjev, ki merijo vitalne znake, in vprašalnika o simptomih, na katerega odgovarja uporabnik. Metoda temelji na algoritmih strojnega učenja na podatkih, ki smo jih zbrali z vprašalniki ter s strokovno zdravniško pomočjo. Diagnostično metodo smo testirali na virtualnih in pravih bolnikih.

Ključne besede
Medicinska diagnostika, vprašalnik

1. UVOD
V zadnjih letih je napredek na področju senzorjev in informacijske tehnologije odprl vrata razvoju naprav in aplikacij s področja medicine, ki bodo namenjene domači uporabi. Osnovna ideja tako imenovanega m-zdravja je, da lahko uporabnik sam spremlja svoje zdravstveno stanje s pomočjo ustreznih naprav. Pri tem gre lahko za preventivni pristop (opozarjanje na morebiten pojav zdravstvenih težav, še preden te postanejo resne) ali za pomoč bolnikom s kroničnimi boleznimi, kot so sladkorna bolezen, kronično srčno popuščanje itd. Pristop m-zdravja koristi tako uporabniku, ki lahko bolje spremlja svoje zdravje, kot tudi zdravstvenemu sistemu, saj omogoča učinkovitejše in hitrejše obravnavanje posameznih bolnikov, krajšanje čakalnih vrst in nižanje stroškov.

Leta 2012 je bil objavljen natečaj Qualcomm Tricorder XPRIZE [1], katerega cilj je razviti napravo, ki bo sposobna spremljati vitalne znake posameznika ter pravilno napovedati serijo različnih bolezni. Na tekmovanju je sodelovala tudi slovenska ekipa MESI Simplifying diagnostics, v okviru katere je skupina z Instituta Jožef Stefan razvila pametno diagnostično metodo [2]. Ta na podlagi meritev vitalnih znakov in vprašalnika o simptomih, na katerega odgovarja uporabnik, določi, kakšen diagnostični test mora uporabnik narediti, da lahko potrdi ali ovrže sum na določeno diagnozo.

V tem prispevku predstavljamo strukturo metode ter trenutne rezultate testiranja na pravih in virtualnih bolnikih. Na koncu predstavimo tudi možne izboljšave pri nadaljnjem delu.

2. METODE
Diagnostična metoda je v obliki aplikacije na voljo uporabniku kot orodje za začetno diagnozo v domači uporabi. Diagnostična metoda temelji na kombinaciji vprašalnika in meritev s senzorji, ki merijo vitalne znake. Primarni vhod za diagnostično metodo (Slika 1) sestoji iz treh različnih vrst podatkov, ki jih aplikacija zajame ob samem začetku diagnostičnega procesa in vključujejo:

(1) razpoznane dejavnike tveganje na podlagi uporabnikovega profila v aplikaciji (npr. prekomerna teža, kajenje, visoka starost itd.),

(2) simptome, razpoznane na podlagi odstopanj izmerjenih vrednosti vitalnih znakov od pričakovanih vrednosti (npr. visok krvni tlak, povišana telesna temperatura itd.), in

(3) lokalizirane bolečinske simptome, ki jih uporabnik označi na anatomskem grafičnem prikazu (npr. glavobol, bolečina v prsih itd.)

Primarni vhodni podatki (dejavniki tveganja in obe vrsti simptomov) se v nadaljnjih korakih diagnostične metode obravnavajo enolično, t.j. kot simptomi. Diagnostična metoda lahko operira izključno s simptomi iz nabora vnaprej definiranih 60 simptomov. Vsak simptom se pri določenem uporabniku obravnava kot *znan* ali *neznan*, pri čemer je vsak simptom, katerega prisotnost je znana, označen kot *prisoten* ali kot *odsoten*.

Iz primarnih vhodnih podatkov diagnostična metoda pridobi informacije (znana prisotnost ali odsotnost) o 37 simptomih. Prisotni simptomi tvorijo (4) začetni nabor prisotnih simptomov.

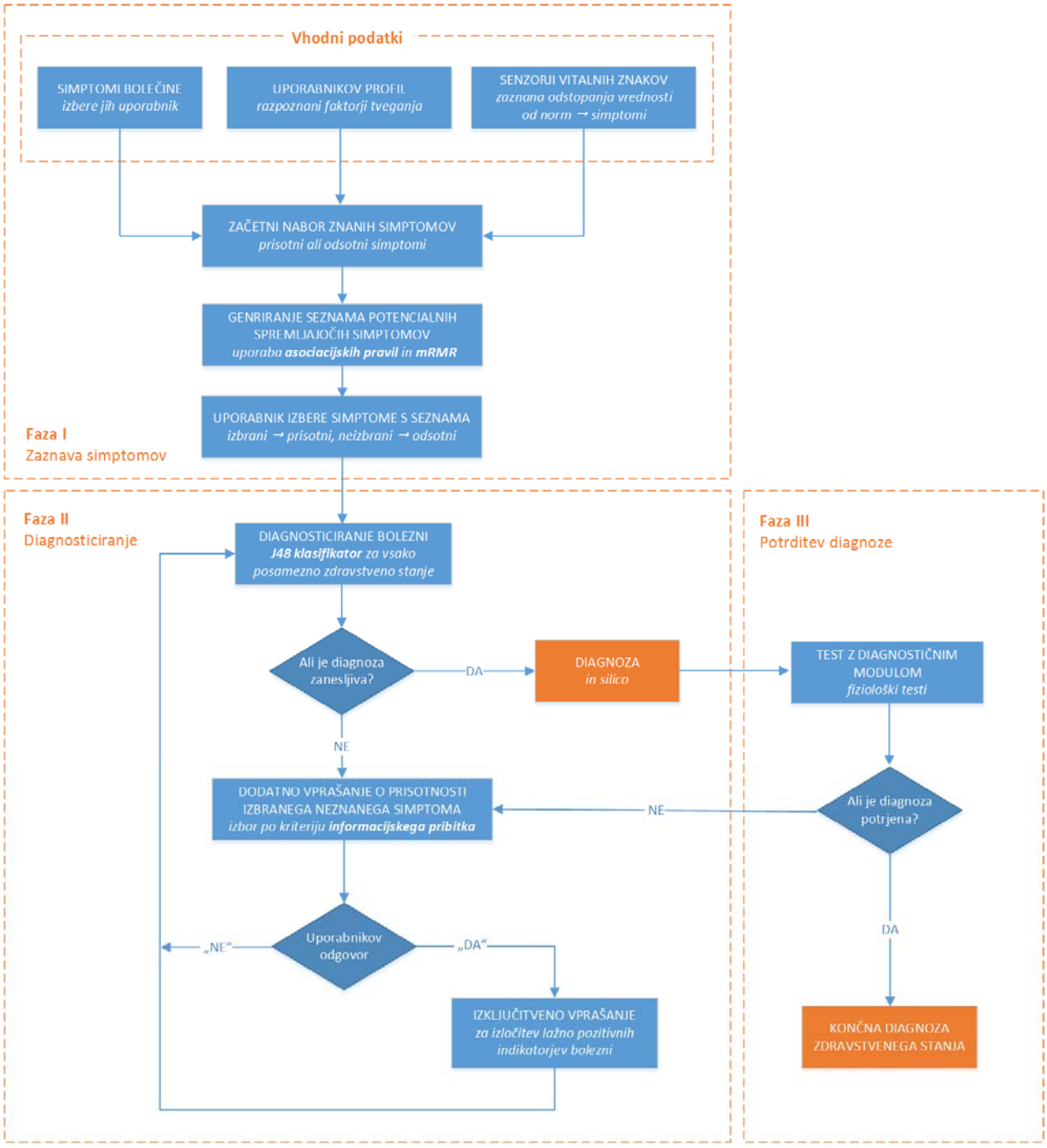
Nabor prisotnih simptomov se v naslednjem koraku diagnostičnega procesa uporabi za generiranje (5) seznama potencialnih spremljajočih simptomov. Na dotični seznam se lahko uvrstijo le simptomi iz množice neznanih simptomov. V kolikor je v začetnem naboru prisotnih simptomov vsaj en simptom, metoda za izbor potencialnih spremljajočih simptomov daje prednost tistim neznanim simptomom, ki se pogosteje (z večjo verjetnostjo) pojavljajo v kombinaciji s simptomi v trenutnem naboru prisotnih simptomov. Pri tem izboru se iščejo asociacijska pravila tipa 'simptom $j \rightarrow$ simptom i' [3–5], kjer je simptom j katerikoli simptom iz nabora prisotnih simptomov in simptom i katerikoli simptom izmed neznanih. Na podlagi pravil z najvišjo zanesljivostjo in zagotovljeno minimalno podporo je med neznanimi simptomi izbrano poljubno število potencialnih spremljajočih simptomov.

Poleg asociacijskih pravil se pri izboru potencialnih simptomov uporablja metoda minimalne redundance in maksimalne relevance (mRMR) [6]. Pri uporabi te metode se izmed neznanih simptomov izbere več simptomov, za katere velja, da:

a) prinesejo največ informacije o končnem razredu, t.j. bolezni (maksimalna relevanca) in

b) so obenem čim manj korelirani z že znanimi simptomi (minimalna redundanca).

Izbor potencialnih simptomov iz nabora neznanih simptomov (Ωs) po metodi mRMR je mogoč tudi v primeru, ko je začetni nabor prisotnih simptomov (S^+) prazen (med znanimi simptomi (S) so torej vsi simptomi odsotni; nabor odsotnih simptomov S^- je v tem primeru enak S).



Slika 1. Diagram poteka celotnega diagnostičnega procesa

Diagnostični proces poteka v treh fazah, pri čemer sta prvi dve fazi, (1) zaznava simptomov in (2) *in silico* diagnosticiranje, del razvite pametne diagnostične metode. V zadnji fazi diagnostičnega procesa poteka, (3) potrditev diagnoze, se izvedejo fiziološki testi z namenskimi diagnostičnimi moduli.

V sklopu pametne diagnostične metode je prisoten tudi korak z izključitvenim vprašanjem. Le-ta je namenjen izključno resničnim uporabnikom (in ne testiranju z virtualnimi), saj je njegov namen zmanjšanje subjektivnosti pri odgovarjanju. Izključitvena vprašanja so za vsak simptom točno določena vnaprej in so zasnovana za prepoznavanje ter izločitev najpogostejših lažnih indikatorjev bolezni. Primer izključitvenega vprašanja za simptom '*kri v urinu*' je *"Ali ste nedavno uživali rdečo peso?"* – saj se zaužita rdeča pesa izloča v urin, pri čemer ga obarva temno rdeče (nepatološko), kar uporabniki pogosto zmotno zamenjujejo s krvjo v urinu (patološko).

Uporabljeno pravilo mRMR za izbor potencialnega simptoma *i* je definirano na podlagi razlike v medsebojni informaciji [6] (alternativno bi bil lahko uporabljen tudi kvocient) po enačbi:

Enačba 1.

$$\boldsymbol{i}, \textbf{kjer}: \max_{\boldsymbol{i} \in \boldsymbol{\Omega s}} [\boldsymbol{I}(\boldsymbol{i}, \boldsymbol{h}) - \frac{\mathbf{1}}{|\boldsymbol{S}|} \sum_{\boldsymbol{j} \in \boldsymbol{S}} \boldsymbol{I}(\boldsymbol{i}, \boldsymbol{j})]$$

Izbira potencialnega simptoma z metodo mRMR se izvede nad naborom neznanih simptomov Ω*s*. Med temi je izbran tisti simptom, za katerega je vrednost funkcije (Enačba 1) največja. Pri tem je *I(i,h)* medsebojna informacija med simptomom *i* iz nabora neznanih simptomov ter boleznijo *h*, končnim razredom. *I(i,j)* je medsebojna informacija med neznanim simptomov *i* in simptomom *j* iz nabora znanih simptomov *S*, ki vsebuje |*S*| simptomov.

Ko je z enačbo mRMR izbran nov potencialni simptom *i*, bo le-ta odstranjen iz množice Ω*s* in dodan v množico potencialnih simptomov *P* (slednja je pred prvo iteracijo metode mRMR prazna). Za izbor še enega ali več dodatnih potencialnih simptomov pa simptoma oz. simptomov v množici *P* ne smemo več obravnavati kot neznane(ga), temveč kot tiste, ki so že (oz. še bodo) znani. Vse nadaljnje iteracije metode mRMR lahko torej opišemo z naslednjo enačbo:

Enačba 2.

$$i, \text{kjer}: \max_{i \in \Omega s} [I(i,h) - \frac{1}{|S| + |P|} \sum_{j \in S \cup P} I(i,j)]$$

Za izbiro vsakega dodatnega potencialnega simptoma se celotna iteracija z metodo mRMR ponovi (Enačba 2) nad naborom preostalih neznanih simptomov Ω*s*, dokler nabor potencialnih simptomov *P* ne doseže želenega števila elementov – simptomov. Zbran nabor potencialnih simptomov *P* se nato v obliki seznama ponudi uporabniku, da označi prisotne simptome (neoznačeni so posledično interpretirani kot odsotni).

V naslednjem koraku se informacije o vseh znanih – tako prisotnih kot tudi odsotnih – simptomih uporabijo za določanje bolezni oz. zdravstvenega stanja (6). Verjetnosti so izračunane za 15 vnaprej definiranih zdravstvenih stanj (14 bolezni in 'zdrav'). Za računanje teh verjetnosti se uporablja množica klasifikatorjev J48, po eden za vsako zdravstveno stanje.

Določena sta dva pragova za srednjo in višjo verjetnost prisotnosti nekega zdravstvenega stanja, ki sta empirično določena na 40% (srednji prag) in 80% (višji prag). V primeru, da verjetnosti vseh zdravstvenih stanj padejo pod srednji prag (zdravstveno stanje je malo verjetno) ali se dvignejo nad višji prag (zdravstveno stanje je zelo verjetno), se diagnoza obravnava kot zanesljiva in *in silico* diagnostični proces se konča (8). V realnem sistemu pri tem ne gre za končno diagnozo, temveč za usmeritev na najprimernejši fiziološki diagnostični test z dodatnimi diagnostičnimi moduli, s katerimi lahko diagnozo dokončno potrdimo ali ovržemo. Če denimo metoda napove, da ima uporabnik krvno bolezen (npr. anemijo), ga napoti na krvni test, v primeru kronične obstruktivne pljučne bolezni pa na test, pri katerem se posluša dihanje.

V primeru, ko se verjetnosti enega ali več zdravstvenih stanj nahajajo v območju med obema pragovoma (40 – 80%), v t.i. sivem območju, se diagnoza obravnava kot nezanesljiva. V tem primeru je potrebna informacija o prisotnosti ali odsotnosti vsaj še enega dodatnega neznanega simptoma. Kot dodatni simptom je med preostalimi neznanimi simptomi izbran tisti simptom *i*, ki ima največji informacijski pribitek *IG(h,i)*, po enačbi:

Enačba 3.

$$\text{IG}(h, i) = \text{H}(h) - \text{H}(h|i)$$

Pri tem je *H(h)* entropija končnega razreda in *H(h|i)* entropija končnega razreda v primeru, da bi bil simptom *i* znan.

Pri naši diagnostični metodi se informacijski pribitek simptomov izračuna na uteženih učnih podatkih, kjer so težje obteženi tisti primeri, katerih razredi sovpadajo s tistimi zdravstvenimi stanji, katerih verjetnosti so v sivem območju. Na ta način izbran simptom je v obliki vprašanja predstavljen uporabniku, ki potrdi njegovo prisotnost ali odsotnost (7).

Pri tem je zaradi uporabniške izkušnje pomembno, da se zanesljiva diagnoza postavi s čim manj dodatnimi vprašanji. Načeloma bi uporabnika lahko povprašali o vseh simptomih s seznama, vendar bi bilo to zamudno in večje število vprašanj nam ne bi nujno prineslo dodatnih informacij o bolezenskem stanju.

Izbor na podlagi informacijskega pribitka omogoča manjše število potrebnih vprašanj, kot če bi bila le-ta fiksno določena vnaprej. Namen uteževanja učnih podatkov je hitrejša konvergenca verjetnosti izven sivega območja in s tem še dodatno zmanjšanje števila potrebnih vprašanj za postavitev zanesljive diagnoze (8).

3. EKSPERIMENTI IN REZULTATI

Z uporabo ekspertnega znanja smo strukturirali tabelo, ki korelira 15 izbranih zdravstvenih stanj s 60 simptomi, določenimi s strani medicinskih strokovnjakov. Tabela je bila uporabljena za generiranje učnih podatkov s 15.000 virtualnimi bolniki in testnih podatkov s 1.500 virtualnimi bolniki. V obeh podatkovnih množicah so bile posamezne bolezni enakomerno zastopane; med učnimi podatki je bilo po 1000 bolnikov z vsako izmed bolezni, med testnimi podatki pa po 100 bolnikov. Celotni podatki so bili uporabljeni za učenje 15 različnih J48 klasifikatorjev, po eden za vsako izmed zdravstvenih stanj.

Testi so pokazali visoko senzitivnost in specifičnost. Na primer, pri 99% bolnikov s hipertenzijo je bolezen tudi ustrezno razpoznana oz. diagnosticirana. Med tistimi, ki pa so diagnosticirani s hipertenzijo, jih 88% tudi zares ima to bolezen. Najslabše je diagnosticiranje zdravega človeka (pri katerem so lahko sicer tudi prisotni simptomi), pri čemer je senzitivnost 61% in specifičnost 62%. Povprečna vrednost senzitivnosti preko vseh zdravstvenih stanj je 88,4%, povprečna specifičnost 88,6% in povprečna točnost 88,3% [7].

4. DISKUSIJA

Rezultati testiranj diagnostične metode na virtualnih bolnikih kažejo visoko senzitivnost, specifičnost in klasifikacijsko točnost (vse nad 80%) diagnostične metode [7]. Vrednosti so najverjetneje preveč optimistične, tudi glede na mnenja strokovnih medicinskih sodelavcev. Glavni izmed razlogov je uporaba ekspertne tabele za generiranje tako učnih kot testnih podatkov. Prav tako je število možnih zdravstvenih stanj zgolj 15, veliko manj kot sicer v praksi. Večina izmed 14 izbranih bolezni se med seboj bistveno razlikuje v simptomih, zaradi česar je med njimi lažje ločiti (višja klasifikacijska točnost). Izjeme so pljučne bolezni (pljučnica, tuberkuloza, kronična obstruktivna pljučna bolezen, spalna apneja), ki so si med seboj tudi bolj podobne po simptomih. V realističnem sistemu je ključno, da diagnostična metoda prepozna, da gre za sum

na pljučno bolezen, saj potem uporabnika napoti na ustrezni test – ta pa nato poda specifično diagnozo.

V nadaljevanju bi bilo zanimivo preučiti, kako se diagnostična metoda obnaša ob dodajanju dodatnih bolezenskih stanj in dodatnih simptomov.

5. LITERATURA

[1] Qualcomm Tricorder XPRIZE, http://www.qualcommtricorderxprize.org/

[2] Somrak M., Gradišek A., Luštrek M., Mlinar A., Sok M., in Gams M.: *Medical diagnostics based on combination of sensor and user-provided data*. (NetMed, ECAI 2014)

[3] McCormick T.H., Rudin C. in Madigan D.: *A Hierarchical Model for Association Rule Mining of Sequential Events: an Approach to Automated Medical Symptom Prediction*. Annals of Applied Statistics. (2012) .

[4] Soni S. in Vyas O.P.: *Fuzzy Weighted Associative Classifier: A Predictive Technique For Health Care Data Mining.* International Journal of Computer Science, Engineering and Information Technology (IJCSEIT), (2012) .

[5] Soni S. in Vyas O.P.: *Using Associative Classifiers for Predictive Analysis in Health Care Data Mining*. International Journal of Computer Applications (2010)

[6] Peng H., Ding C.: *Minimum Redundancy Feature Selection from Microarray Gene Expression Data.* Journal of Bioinformatics and Computational Biology (2005)

[7] Somrak M., Luštrek M., Sušterič J., Krivc T., Mlinar A., Travnik T., Stepan L., Mavsar M., Gams M.*: Tricorder: Consumer Medical Device for Discovering Common Medical Conditions*. Informatica 38, Ljubljana (2014)