Zbornik radova Fakulteta tehničkih nauka, Novi Sad

UDK: 004.7
DOI: https://doi.org/10.24867/17BE02Krajcinovic

SISTEM ZA OBRADU PODATAKA PAMETNOG DOMA

SMART HOME DATA PROCESSING SYSTEM

Nemanja Krajčinović, Dejan Nemec, *Fakultet tehničkih nauka, Novi Sad*

Oblast – ELEKTROTEHNIKA I RAČUNARSTVO

Kratak sadržaj – *Cilj ovog rada jeste da razvije sistem za obradu velikih količina podataka iz okruženja pametnog doma u cilju izvođenja zaključaka i omogućavanja predviđanja različitih događaja koji pospešuju rad pametnog doma i čine svakodnevni život lakšim.*

Ključne reči: *internet stvari, pametni dom, obrada velikih količina podataka.*

Abstract – *This paper aims to develop a system for processing large amounts of data from the smart home environment to draw conclusions and enable the prediction of various events that enhance the work of the smart home and make everyday life easier.*

Keywords: *internet of things, smart home, big data.*

1. UVOD

Računarska tehnika se poslednjih 30 godina razvija u mnogo različitih smerova ali sa jednim ciljem – da poboljša život ljudi. Određene oblasti su već dovedene na visok nivo apstrakcije i kao takve, pomažu u svakodnevnim aktivnostima i olakšavaju život. Naravno, to nije razlog za stagniranje trenda razvoja tih oblasti, kao i otkrivanja novih. Početak 21. veka doneo je nagli razvoj fizičke arhitekture i programske podrške računara što je omogućilo i razvoj mašinskog učenja. Danas, mašinsko učenje se koristi u različitim oblicima na velikom broju mesta i prisutno je čak i bez znanja korisnika.

Ovaj rad predstavlja jedan od primera gde je upotreba različitih tehnika optimizacije velike količine podataka i mašinskog učenja iskorišćena u praktične svrhe. U sprezi sa tehnologijama iz oblasti računarskih komunikacija i fizičke arhitekture računara, kao i programske podrške, cilj ovog rada bio je da se optimizuje baza podataka nastala u jednom privatnom stambenom prostoru i na osnovu te baze kreiranje predikcionih modela za potrošnju električne energije.

U drugom poglavlju iznete su osnove i aspekti koje predstavljaju potporu za ovaj rad. U trećem poglavlju iznet je koncept celokupnog sistema, kao i neki od ključnih aspekata zašto je određena tehnologija izabrana. Detaljan prikaz algoritma predstavljen je u četvrtom poglavlju, kao i specifičnosti koje su se pojavile pri izradi zadatka a takođe i kako su prevaziđene. Postupak testiranja i ostvareni rezultati prikazani su u petom poglavlju, a zaključak u šestom.

NAPOMENA:
Ovaj rad proistekao je iz master rada čiji mentor je bio dr Živko Bojović, vanr. prof.

2. OSNOVE RADA

2.1. Internet stvari

U poslednjoj deceniji, kućni aparati prešli su put od izrade jednostavnih uređaja koji su podržani jeftinim senzorima, do inteligentnih uređaja koji mogu detektovati kretanje čoveka. Pametna kuća predstavlja okruženje u kome se kućni aparati mogu nadgledati i kontrolisati daljinski, na osnovu različitih ugrađenih uređaja kao što su senzori i aktuatori. Neki sistemi pametnih kuća, koji su razvijeni poslednjih godina, mogu da detektuju prisutnost na osnovu senzora pokreta. Pasivni senzori, zasnovani su na infracrvenim i drugim sličnim tehnologijama, i u tu grupu spadaju i senzori pokreta. Nedostatak tih tehnologija je što ne detektuju objekte kada korisnik nije dugo u pokretu. Alternativa za prevazilaženje ovog nedostatka je upotreba neprekidnih tehnologija zasnovanih na kompjuterskoj viziji ili drugih sličnih sveprisutnih tehnologija detekcije objekata. Aktivni senzori prikupljaju značajne informacije, uključujući privatne podatke korisnika, tako da je izazov ispravno odvojiti anonimne podatke. Postoje dva glavna razloga za sprečavanje usvajanja aktivnih i invazivnih senzoriskih tehnologija u pametnim kućama:

- teško generalizovanje modela sa promenljivim uslovima okruženja,
- briga korisnika za uočenu nametljivost.

Ponašanje senzora i aktuatora (nametljivo ili nenametljivo), može značajno varirati u zavisnosti od konteksta objekta (kuće, zgrade). To je moguće precizno postići praćenjem zauzetosti pomoću dostupnih informacija iz nenametljivih izvora podataka kao što su sobna temperatura, rashladni uređaji, potrošnja vode i niz drugih povezanih uređaja [1]-[3].

Predmet ovog istraživanja je inteligentno izračunavanje na osnovu preferencija korisnika pomoću tehnologije velikih količina podataka prikupljenih u IoT okruženju korišćenjem nenametljivog pristupa modelovanja. Uključivanjem personalizacije zasnovane na obrascima ponašanja korisnika IoT okruženja, predloženi model sistema može pružiti pametniji odgovor bez potrebe za intervencijom korisnika, čime postiže nenametljivu automatizaciju pametne kuće.

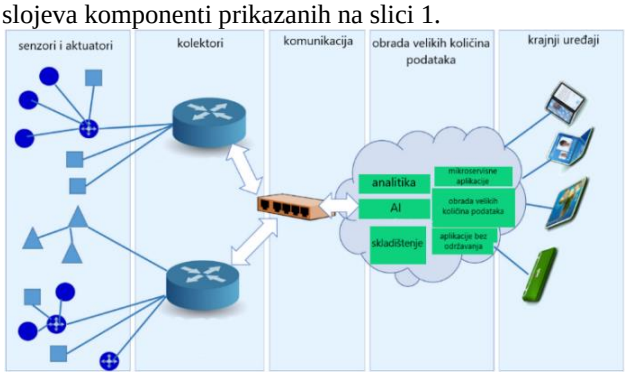
2.2. Obrada velikih količina podataka

Identifikovanje obrazaca ljudskog ponašanja u realnom vremenu povezanih sa pametnom kućom prilika je za modelovanje personalizacije. To se može postići korisničkim profilisanjem podataka u realnom vremenu generisanim sa uređaja i senzora kojima upravlja korisnik.

Osim toga, različiti elementi podataka i atributi iz različitih izvora životne sredine mogu se analizirati kako

bi se razumeli obrasci podataka koji prikazuju ponašanje određenog korisnika i fizičke aktivnosti koje omogućavaju personalizovano profilisanje korisnika u pametnom domu [4]-[6].

Generička IoT arhitektura koja proširuje mogućnosti obrade velikih količina podataka obično se sastoji od pet slojeva komponenti prikazanih na slici 1.



Slika 1. *Generička IoT arhitektura*

Predloženi model za automatizaciju pametnih kuća zasnovan je na masivnom nenametljivom prikupljanju podataka i generalnom konsenzusu tehnologije velikih količina podataka – četiri V (raznolikost, verodostojnost, brzina, vrednost).

2.3. Mašinsko učenje

Za razvoj prototipa prilagođavaju se parametri pametnog doma, poput temperature i osvetljenja, koristeći regresore. Model namerava da predvidi takve obrasce tokom realnog vremena, tako da se izvrše neophodna prilagođavanja parametara [7].

3. KONCEPT REŠENJA

Prototip rešenja pretpostavljenog sistema delom je razvijen i implementiran u jednom privatnom okruženju. Vodeći računa o aspektima nenametljivosti, senzori i aktuatori su postavljeni prema smernicama iz prethodnog poglavlja. Implementirani sistem obuhvata širok spektar senzora za različite uređaje, od kuhinjskih električnih uređaja, preko prekidača, utičnica, bojlera, pa sve do informativnih stanja mobilnih uređaja koji su se nalazili u domu.

Podaci su prikupljani 4 meseca i to u periodu od oktobra 2020. godine do februara 2021. godine. Beleženi su u pravilnim vremenskim intervalima – svakog minuta, gde je u svakom minutu zapisivano stanje sa senzora iz okruženja. Takvo okruženje generiše veliku količinu podataka sa mnoštva različitih uređaja, koji se skladište na kolektoru. Ovakvo rešenje pripada konvencionalnim IoT implementacijama pametnog doma, koja se danas u praksi često mogu sresti. Ono što ih karakteriše je nemogućnost predviđanja događaja u budućnosti i automatizacija upravljačkih funkcija. Osnovna ideja ovog sistema koje predstavlja fundament istraživanja je mogućnost prevazilaženja ovih problema, upravo primenom veštačke inteligencije, odnosno tehnologija mašinskog učenja i obrade velikih količina podataka.

Prilikom implementacije predloženog rešenja najveći značaj je pridat ispunjavanju konsenzusa o nenametljivosti sistema pametnog doma, ispunjavanju svih standarda po pitanju komunikacija između uređaja i kolektoru sistema. Međutim, nedostatak ovog sistema je nepostojanje modela efikasnog skladištenja takvih, velikih količina podataka, kao i njihove obrade. Zato ovaj rad predstavlja koncept i primer implementacije takvog sistema, koji bi omogućio efikasno skladištenje podataka, njihovo prečišćavanje i obogaćivanje, kao i implementaciju modela estimacije ponašanja korisnika i samim tim i personalizacije sistema.

Prečišćavanje i obogaćivanje podataka uvedeno je u cilju da se optimizuje skladišni prostor i obezbedi njegovo efikasno korišćenje. Ovaj proces je izvršen primenom statističkih analiza i izvođenjem odgovarajućih zaključaka na osnovu njih. Zato i najveći doprinos ovog istraživanja, upravo leži u razvoju i implementaciji ovog dela sistema koji obezbeđuje nenametljivu automatizaciju upravljačkih funkcija pametnog doma.

Nakon sređivanja baza podataka u skladištima, primenjeno je mašinsko učenje kao tehnologija, odnosno ključni aspekt koji može obezbediti predikciju ponašanja korisnika sistema, kao i personalizaciju istog. U radu su date određene pretpostavke, modeli koji su isprobani i opisani rezultati njihove primene, koji su potom upoređeni i objašnjeni.

4. PROGRAMSKO REŠENJE

Implementirani sistem ne poseduje potrebne mehanizme za efikasno skladištenje i obradu podataka. Baza podataka je na početku imala 176.830 uzoraka i 250 obeležja.

Raščlanjivanje podataka je prvi deo sistema gde će ovakvi, nestruktuirani podaci, biti prevedeni u struktuirane podatke, u odgovarajućim formatima. Algoritam obrade jedne log datoteke se sastoji iz sledećih koraka:

1. Učitavanje jedne linije te datoteke.
2. Prevođenje ključa linije u format datuma i vremena kada je linija zabeležena.
3. Pribavljanje imena svakog senzora, kao i vrednosti koja je zabeležena o tom senzoru.
4. Zapis takve linije kao jednog uzorka u CSV bazu podataka.

Nakon što su podaci prevedeni u jedinstvenu CSV bazu podataka, izvršeno je prevođenje niza karaktera koji predstavljaju datum i vreme, u pogodan oblik za dalju analizu, a to je poseban zapis vremena, dana, meseca i godine za svaki uzorak iz baze podataka.

U cilju uklanjanja redundantnih podataka prvo se vrši provera broja nedostajućih vrednosti po obeležju. Algoritam je sledeći:

1. sumiranje praznih polja po kolonama,
2. zapis konačnih vrednosti u listu sa imenima obeležja.

Na osnovu ovog algoritma izvučen je jedan veoma zanimljiv zaključak: čak 26 obeležja u ovoj bazi imaju preko 99% nedostajućih podataka, što automatski dovodi do akcije uklanjanja tih obeležja iz baze podataka.

Vizuelnim analiziranjem i testiranjem baze zaključeno je da postoji još jedan problem, a to je nepravilan zapis različitih stanja na nivou cele baze, odnosno, korišćenje različitih ključnih reči i oznaka za jednu te istu vrednost. Tako su nedostajuće vrednosti označavane sa „*None*“, „*unavailable*“, „*unknown*“, „*disarmed*“. Nakon toga, uzorci koji sadrže nedostajuće podatke su uklonjeni iz baze. Uklonjeni su uzorci sa barem jednim nedostajućim podatkom i broj uzoraka je umanjen na 151.779 uzoraka.

Sledeće je trebalo definisati tipove obeležja, kako bi podaci bili iskoristljivi u daljem procesu. Definisanje tipa svakog obeležja je rađeno na osnovu analize histograma i utvrđivanja kakve se vrednosti pojavljuju na nivou jednog obeležja. Ovim postupkom je zaključeno da je 95 obeležja numeričko. Dalje, 92 obeležja su identifikovana kao kategorička obeležja, sa binarnim ishodima.

Analizom je dodatno utvrđeno da postoje obeležja sa tekstualnim informacijama koja u ovakvom sistemu ne mogu biti od koristi. Tim postupkom je uklonjeno 26 obeležja. Konačno, baza je prevedena u pogodan format za dalju upotrebu sa 151.779 uzoraka i 191 obeležjem.

Prečišćavanje i obogaćivanje podataka je proces kojim se može obezbediti redukcija podataka bez gubitka informacija koji se nalaze u podacima.

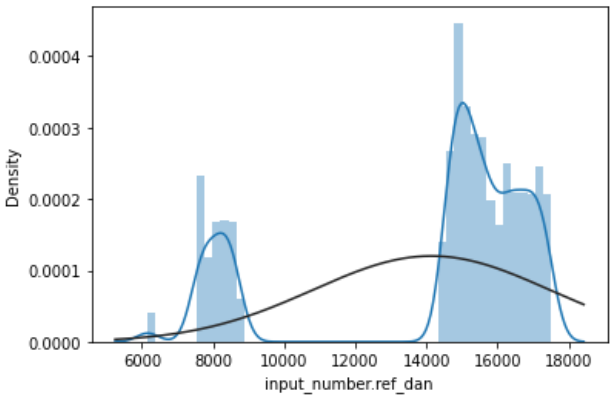
Prvi korak je analiziranje histograma pojavljivanja vrednosti unutar svakog obeležja i raspodele tih vrednosti. Prilikom vizuelne analize rezultata zapažena je jedna nepravilnost. Veliki broj obeležja imao je raspodelu vrednosti visoko koncentrisanu oko jedne vrednosti, što se moglo videti i na histogramu pojavljivanja vrednosti za to obeležje. Na histogramu je bilo vidljivo i da se jedna vrednost pojavljuje veliki broj puta, skoro pa približan broju uzoraka. Zbog toga je sproveden algoritam za uklanjanje obeležja kod kojih se jedna vrednost ponavlja i to u više od 99% u odnosu na ukupan broj uzoraka. Opravdanje za uklanjanje ovakvih obeležja leži u činjenici da takva obeležja ne nose informaciju za ovaj sistem i samim tim nikako ne mogu doprineti razvoju ovog sistema. Ovim algoritmom je uočeno i uklonjeno čak 49 obeležja iz baze podataka.

Drugi korak ovog dela analiziranja baze podataka bio je kreiranje i analiza korelacione matrice.

S obzirom da baza podataka i dalje sadrži veliki broj obeležja, nije bilo moguće jasno prikazati matricu sa vidljivim odnosima između obeležja, ali se uočilo da je u nekim delovima korelacija približna maksimumu, što znači da postoje obeležja koja su visoko korelisana i potrebno ih je ukloniti. Detaljnijom analizom ove matrice, utvrđeno je da se visoke korelacije javljaju između obeležja čiji je sufiks imena sličan ili se može pretpostaviti da se odnose na istu ili sličnu stvar, a da se prefiksi razlikuju. Odatle se može izvući sledeći zaključak: postoje obeležja koja su izvedena iz osnovnog skupa obeležja ovog sistema. Detaljnom analizom korelacione matrice utvrđeno je da sva obeležja koja imaju korelacioni koeficijent preko 0,6 sa nekim drugim obeležjem mogu biti uklonjena iz baze podataka. Algoritam uklanjanja suvišnih obeležja izveden je algoritmom korišćenja korelacione matrice i to podataka ispod glavne dijagonale, gde su, idući red po red, odnosno, obeležje po obeležje, uklanjanja ona koja imaju korelacioni koeficijent veći od 0,6 sa analiziranim obeležjem. Ovim postupkom uklonjeno je 70 obeležja iz baze podataka, čime je baza podataka redukovana na 72 obeležja.

Treći korak ovog dela predstavlja analiza ciljnog obeležja za predikciju. Informacija koja je dobijena uz bazu podataka jeste da je obeležje *input_number.ref_dan* ciljano obeležje, da ono predstavlja zbirnu potrošnju električne energije doma na nivou sve tri faze i da je izraženo u kilovat-časovima (kWh). Na osnovu tog podatka, na slici 2 prikazan je histogram pojavljivanja vrednosti unutar tog obeležja u odnosu na normalnu raspodelu.

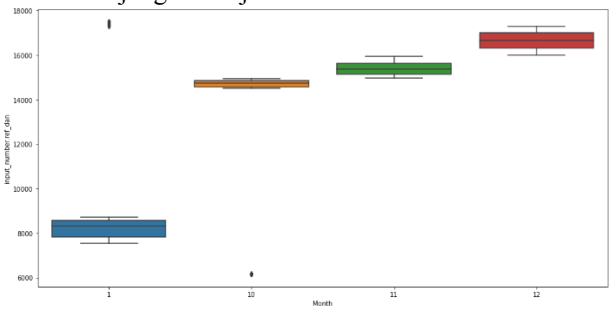
Na osnovu ovog histograma može se jasno uočiti da se ovde vrednosti grupišu oko vrednosti oko 8.000 i 16.000 i da se ne može estimirati normalnom raspodelom na nivou celog uzorka ovog obeležja.



Slika 2. *Histogram ciljnog obeležja sa estimacijom u odnosu na normalnu raspodelu*

Takođe, ciljno obeležje je statistički analizirano i u odnosu na vremenske podatke.

Na slici 3 može se videti da su raspodele poređane u rastućem poretku u odnosu na mesec, a isti zaključci se mogu uvažiti i za raspodele vrednosti u odnosu na dane u mesecu i vreme u toku jednog dana. Na osnovu tih grafika može se izvući konačan zaključak, koji je potvrđen i vizuelnom analizom ovog obeležja: ciljno obeležje je zapisivano kumulativno, odnosno, vrednosti su akumulisane i zbog toga se na graficima pojavljuju rastući poreci. Takođe, krajem januara 2021. godine vrednost je umanjena i zato se desila promena pika podataka u poslednjem delu baze podataka ciljnog obeležja.



Slika 3. *Raspodela uzoraka ciljnog obeležja u odnosu na mesece u godini*

Automatizacija pametnog doma i personalizacija takvog sistema leže u primeni veštačke inteligencije na ovakve podatke. Ideja ovakvog sistema svakako nije predviđanje samo jednog obeležja, potrošnje električne energije na nivou celog doma, ali zbog nedostatka informacija o obeležjima, ovde će biti isprobano samo predviđanje potrošnje električne energije.

Baza podataka je prvobitno podeljena na trening i test skup u odnosu 9:1 i to uz prvobitno mešanje uzoraka na slučaj. Modeli koji su testirani jesu:

- opšti model linearne regresije,
- model linearne regresije sa hipotezom koja sadrži interakcije između obeležja,

- opšti model linearne regresije sa standardizovanim obeležjima,
- *Ridge* regresioni model,
- *Lasso* regresioni model.

Kao kriterijum pronalaženja optimalnog modela linearne regresije korišćen je RSS. U tabeli 1 predstavljene su RSS vrednosti svakog testiranog modela.

Tabela 1. *Rezultati obuke različitih regresionih modela*

Regresioni model	RSS
Opšti model linearne regresije	35.698.190.141
Model linearne regresije sa hipotezom koja sadrži interakcije između obeležja	2.575.054.607
Opšti model linearne regresije sa standardizovanim obeležjima	35.698.190.141
Ridge regresioni model	35.698.232.742
Lasso regresioni model	35.697.059.689

Iz tabele rezultata se može zaključiti da su rezultati modela daleko od prihvatljivih rezultata i da ovakvi modeli nisu iskoristivi. Dodatnom analizom su utvrđene dve nepravilnosti u okviru ciljnog obeležja, a to su:

- rezultati potrošnje električne energije su zapisivani kumulativno, gde početak nije jasno označen,
- postoje dva različita pika u podacima i to je uočeno da se dešava zato što je u jednom trenutku, sredinom jaunara 2021. godine, rezultat umanjen za neku vrednost, što ovde nije objašnjivo niti shvatljivo zašto se desilo.

Ove dve nepravilnosti mogle su biti primećene na slici 2. Zbog ovih problema pokušano je da se određenim algoritmima prevaziđu ovi problemi, ali se pokazalo da postoje još neki dodatni problemi zbog kojih nije moguće jednoznačno ukloniti navedene nepravilnosti.

5. REZULTATI

U trećem poglavlju objašnjeno je da je sistem pametnog doma kreiran u privatnom vlasništvu zadržao konstrukciju kolektora i skladišta na nivou jednog uređaja i nije brinuo o efikasnom skladištenju podataka, kao ni o mogućoj obradi tih podataka. Zato su podaci zapisivani u tekstualnom formatu u *log* fajlove za 4 meseca zauzimali 4,35 GB. Raščlanjivanjem JSON objekata i pakovanjem stanja senzora u CSV bazu, veličina je redukovana na 232,6 MB. Sprovedenim operacijama za uklanjanje neiskoristljivih obeležja, kao i redefinisanjem tipova obeležja u odgovarajući oblik veličina baze podataka redukovana je na 96,94 MB. Redukcijom obeležja sa visokim korelacionim koeficijentom baza je redukovana na veličinu od 43,13 MB. Uzevši u obzir da je originalna veličina podataka bila 4,35 GB, faktor kompresije iznosi 103,58, što znači da je ostvarena ušteda skladišnog prostora 99,03%.

Najbolji model, model linearne regresije sa hipotezom koja sadrži interakcije između obeležja evaluiran je pomoću srednje apsolutne greške (MAE), kroz unakrsnu validaciju, gde je baza podeljena na 10 podskupova sa odgovarajućim parametrima i srednja apsolutna greška je 3315,86. Vizuelnom analizom ciljnog obeležja, zaključeno je da su promene između kumulativnih beleženih vrednosti reda 10^{-2}, a negde čak i 10^{-3}. Iz ovoga se takođe može zaključiti da ovakva srednja apsolutna greška predstavlja veoma lošu mogućnost za predviđanje potrošnje električne energije.

6. ZAKLJUČAK

U ovom radu prikazano je kako se različite tehnologije danas dostupne mogu praktično upotrebiti za dobijanje informacija koje čovek ne može da pribavi, ili može ali na vrlo težak način. Iako dobijeni rezultati nisu na visokom nivou, uspešno je postignuto izvođenje zaključaka šta je potrebno unaprediti i promeniti na sistemu pametnog doma, kako bi se podaci mogli upotrebiti kao kvalitativni podaci u metodama mašinskog učenja.

7. LITERATURA

[1] Stojkoska, B.L.R.; Trivodaliev, K.V. A review of Internet of Things for smart home: Challenges and solutions. *J. Clean. Prod.* 2017, *140*, 1454–1464,

[2] Choi, D.; Choi, H.; Shon, D. Future changes to smart home based on AAL healthcare service. *J. Asian Arch. Build. Eng.* 2019, *18*, 190–199,

[3] Helal, S.; Bull, C.N. From Smart Homes to Smart-Ready Homes and Communities. *Dement. Geriatr. Cogn. Disord.* 2019, 47, 157–163,

[4] Benmansour, A.; Bouchachia, A.; Feham, M. Multioccupant Activity Recognition in Pervasive Smart Home Environments. *ACM Comput. Surv.* 2016, *48*, 1–36,

[5] Russell, L.; Goubran, R.; Kwamena, F. Personalization Using Sensors for Preliminary Human Detection in an IoT Environment. In Proceedings of the 2015 International Conference on Distributed Computing in Sensor Systems, Fortaleza, Brazil, 10–12 June 2015; pp. 236–241,

[6] Li, R.Y.M.; Li, H.C.Y.; Mak, C.K.; Tang, T.B. Sustainable Smart Home and Home Automation: Big Data Analytics Approach. *Int. J.Smart Home* 2016, 10, 177–198,

[7] Overmars, A.; Venkatraman, S. Towards a Secure and Scalable IoT Infrastructure: A Pilot Deployment for a Smart Water Monitoring System. *Technologies* 2020, *8*, 50.

Kratka biografija:

Nemanja Krajčinović rođen je u Šapcu 1998. godine. Osnovne akademske studije, na smeru Komunikacione tehnologije i obrada signala, završio je 2020. godine, sa odbranom bečelor rada pod naslovom „Praćenje kretanja i zadržavanja ljudi u zatvorenom prostoru primenom kompjuterske vizije". Do sada je je objavio 1 naučni rad na međunarodnom skupu TELFOR.

kontakt: kr.nemanjja@gmail.com

Dejan Nemec rođen je 1972. god. Diplomirao, specijalizirao i magistrirao je na Fakultetu tehničkih nauka iz oblasti Elektrotehnike i računarstva. Oblast interesovanja su telekomunikacije i obrada signala.

Zahvalnica:
Izradu ovog rada pomogao je Fakultet tehničkih nauka u Novom Sadu, Departman za energetiku elektroniku i telekomunikacije, u okviru projekta pod nazivom „Razvoj i primena savremenih metoda u nastavi i istraživačkim aktivnostima na Departmanu za energetiku, elektroniku i telekomunikacije".