

НАСТАВНО-НАУЧНОМ ВЕЋУ

Предмет: Подобност теме и кандидаткиње Марије Костић, мастер инжењера електротехнике и рачунарства и мастер инжењера информационих технологија, за израду докторске дисертације.

Одлуком Наставно-научног већа Електротехничког факултета бр. 717/29 од 16. маја 2023. године, именовани смо за чланове Комисије за оцену подобности теме и кандидаткиње Марије Костић, мастер инжењера електротехнике и рачунарства и мастер инжењера информационих технологија, за израду докторске дисертације и научне заснованости теме под називом: „**Интелигентни систем за семантичку претрагу програмског кода**“ (енгл. „*Intelligent system for semantic search of source code*“).

На основу материјала приложеног уз Захтев кандидата, Комисија подноси следећи

ИЗВЕШТАЈ

1. Подаци о кандидату

1.1. Биографски подаци

Марија Костић, мастер инжењер електротехнике и рачунарства и информационих технологија, рођена је 23.05.1997. године у Београду. Основну школу „Аца Милосављевић“ у Београду завршила је као носилац Вукове дипломе и Ђак генерације. Након тога је уписала Прву спортску кошаркашку гимназију у Београду коју је завршила такође као Ђак генерације.

Основне студије на Електротехничком факултету у Београду уписала је 2015. године на студијском програму Софтверско инжењерство. Дипломирала је 2019. године са просечном оценом 9,96 као најбољи студент на студијском програму. Дипломски рад на тему „Софтверски систем за визуелизацију података коришћењем мапа“ под менторством доц. др Дражена Драшковића одбранила је у септембру 2019. године са оценом 10. Током основних студија била је ангажована као студент демонстратор на Катедри за рачунарску технику и информатику. Добитница је студентске стипендије и стипендије за изузетно надарене ученике и студенте Министарства просвете, као и награде задужбине „Гордана Јокић Кашиковић и Драгиша Кашиковић“. Два пута је представљала факултет на Међународном сусрету студената електротехнике – Електријади.

Мастер академске студије на Електротехничком факултету у Београду је уписала октобра 2019. године на модулу Софтверско инжењерство. Положила је све испите са просечном оценом 10,0. Мастер рад на тему „Примена техника машинског учења и анализе социјалних мрежа у предвиђању исхода тениских мечева“ под менторством доц. др Дражена Драшковића одбранила је у септембру 2020. године са оценом 10. Такође је у октобру 2019.

године уписала мастер академске студије у оквиру Мастер 4.0 програма Напредне информационе технологије у дигиталној трансформацији на модулу Анализа података као студент прве генерације. Положила је све испите са просечном оценом 9.9. Мастер рад на тему „Једнојезична, вишејезична и међујезична класификација коментара из програмског кода“ под менторством др Бошка Николића обрала је у мају 2023. године са оценом 10.

Током студије је обавила више стручних пракси као софтверски инжењер, и то: 2019. године у Развојном центру компаније Мајкрософт у Београду и у компанији Гугл у Цириху, 2021. године у компанији Гугл у Амстердаму и 2022. године у компанији Квадратур у Лондону и компанији Гугл у Цириху.

Докторске академске студије на Електротехничком факултету у Београду, на модулу Софтверско инжењерство, је уписала 2020. године. Положила је све испите са оценом 10. Област њеног научног рада је примена обраде природних језика у решавању проблема у софтверском инжењерству. У звање истраживач-приправник изабрана је 21. јануара 2021. године. Од марта 2021. је запослена у Иновационом центру Електротехничког факултета у Београду као истраживач. Учествовала је у пројекту „Нова решења у развоју софтвера заснована на сличности текста“ (AVANTES) који је финансиран од стране Фонда за науку Републике Србије. Ауторка је више радова на међународним конференцијама и у научним часописима.

1.2. Стечено научноистраживачко искуство

Марија Костић је уписала докторске студије на Електротехничком факултету у Београду школске 2020/2021. године на модулу Софтверско инжењерство. Током докторских студија је испунила све обавезе и положила све испите предвиђене наставним планом и програмом студија са просечном оценом 10. Списак положених предмета и обавеза је дат у наредној табели:

Година	Назив предмета/обавезе	Оцена	ЕСПБ
1.	Принципи програмских језика	10 (десет)	9
1.	Интернет програмирање	10 (десет)	9
1.	Напредни алгоритми машинског учења	10 (десет)	9
1.	Вештачка интелигенција и експертски системи	10 (десет)	9
1.	Студијски истраживачки рад у вези докторске тезе I	/	24
2.	Софтверска техника	10 (десет)	9
2.	Неуралне мреже	10 (десет)	9
2.	Увод у научни рад	10 (десет)	6
2.	Научно стручни рад	/	3
2.	Студијски истраживачки рад у вези докторске тезе II	/	24

Током школске 2021/2022. године кандидаткиња је учествовала у реализацији предмета „Управљање софтверским пројектима“ који се одржава на Електротехничком факултету Универзитета у Београду на трећој години основних академских студија (одсек Софтверско инжењерство и модул Рачунарска техника и информатика). Такође је учествовала у пројекту „Нова решења у развоју софтвера заснована на сличности текста“ (AVANTES) који је финансиран од стране Фонда за науку Републике Србије.

Кандидаткиња је аутор 5 научних радова, од којих је један прихваћен за објављивање у врхунском међународном часопису са импакт фактором, а четири рада су објављена у зборницима радова међународних скупова. Списак научних публикација дат је у наставку:

Категорија M20: Научни часописи међународног значаја

- [1] **Kostić M.**, Batanović V., Nikolić B.: *Monolingual, Multilingual and Cross-lingual Code Comment Classification*, - Engineering Applications of Artificial Intelligence, ISSN: 0952-1976, прихваћен за публикување. [M21, IF=7,802]

Категорија M30: Зборници међународних научних скупова

- [2] **Kostić M.**, Srbijanović A., Batanović V., Nikolić B.: *Code Comment Classification Taxonomies*, Proceedings of the 9th International Conference on Electrical, Electronic and Computing Engineering, IcETRAN 2022, Novi Pazar 2022., pp. VII1.1, ISBN: 978-86-7466-930-3. [M33]
- [3] Obradović M., **Kostić M.**, Knežević B., Drašković D.: *An overview of software code review tools and the possibility of their application in teaching at the School of Electrical Engineering in Belgrade*, Proceedings of the 9th International Conference on Electrical, Electronic and Computing Engineering, IcETRAN 2022, Novi Pazar 2022., pp. SS-EDU1.4, ISBN: 978-86-7466-930-3. [M33]
- [4] **Kostić M.**, Petrović M., Drašković D.: *Deep Convolutional Neural Networks for COVID-19 detection from CT scans: A survey*, Proceedings of the 11th International Conference on Information Society and Technology (ICIST), Kopaonik 2021, pp. 143-148, ISSN: 2738-1447. [M33]
- [5] **Kostić M.**, Drašković D.: *Complex Network Analysis of Women's Singles Tennis Matches*, Proceedings of the 28th Telecommunications Forum (TELFOR), Beograd 2020, pp. 1-4, ISBN: 978-166540499-0, doi: [10.1109/TELFOR51502.2020.9306598](https://doi.org/10.1109/TELFOR51502.2020.9306598). [M33]

Дана 09.06.2023. године на Електротехничком факултету у Београду, кандидаткиња Марија Костић је имала јавни испит о подобности предложене теме докторске дисертације и кандидаткиње, пред Комисијом у саставу:

1. др Захарије Радивојевић, ванредни професор, Универзитет у Београду – Електротехнички факултет (председник комисије),
2. др Борис Делибашевић, редовни професор, Универзитет у Београду – Факултет организационих наука,
3. др Татјана Лутовац, редовни професор, Универзитет у Београду – Електротехнички факултет,
4. др Вук Батановић, научни сарадник, Иновациони центар Електротехничког факултета Универзитета у Београду.
5. др Дражен Драшковић, доцент, Универзитет у Београду – Електротехнички факултет,

Комисија је једногласно закључила да кандидаткиња испуњава услове за рад на предложеној теми докторске дисертације и оценила испит оценом – **задовољно**.

1.3. Оцена подобности кандидата за рад на предложеној теми

На основу претходног приказаног резултата досадашњег рада и активности Марије Костић и узимајући у обзир претходно наведене чињенице, Комисија је стекла увид у способности и научно-истраживачки рад кандидаткиње, те констатује да је кандидаткиња стекла неопходно искуство за научно-истраживачки рад. Комисија сматра да је кандидаткиња способна да током израде докторске дисертације оствари резултате, који могу бити приказани и верификовани од стране међународне научне заједнице. Комисија даје позитиван став да кандидаткиња Марија Костић настави са својим истраживачким радом и констатује да кандидаткиња испуњава све потребне и формалне услове да приступи изради дисертације.

2. Предмет и циљ истраживања

Данас се у дигиталним системима чувају масивне количине података. Да би људи могли брзо и једноставно да пронађу тражене информације, развијени су моћни претраживачи. Тренутно милијарде људи свакодневно користи комерцијалне претраживаче за проналажење различитих врста информација као што су текстови, слике или видео записи [1].

Како софтвер постаје све значајнији у различитим аспектима наших живота, програмски код се производи у невероватно великим количинама. Само на популарној платформи отвореног кода *GitHub* је током 2022. године креирано више од 85 милиона нових пројеката [2], а један комплексан софтверски пројекат, као што је језгро Линукс оперативног система може да садржи више милиона редова кода. Због тога је већина кода који неки програмер треба да напише већ написан негде другде или је барем слична постојећем коду [3, 4]. Како би имали користи од ове појаве, програмери често морају да претражују пројекте како би пронашли код који им је релевантан у датом тренутку. На пример, програмер претражује репозиторијум на коме ради како би проверио да ли је и на који начин нека функционалност већ имплементирана, затим да би схватио како се неки код користи или да би пронашао сва појављивања кода који мора да промени ради решавања неког бага [5]. Свакако поред самог кода на коме ради, програмери претражују и друге пројекте у оквиру своје организације, али и пројекте отвореног кода. Све ове и сличне активности називају се *претрагом кода*.

С обзиром на то да је претрага кода веома честа и битна активност у оквиру софтверског инжењерства, постала је популарна истраживачка тема те су предложени многи приступи за њену реализацију. Прва решења су била базирана на техникама претраживања информација (енгл. *Information retrieval*) и третирали су програмски код као текстуалне фрагменте природног језика [6, 7, 8]. Међутим, фундаментални проблем ових решења је неусклађеност између намере програмера која је исказана упитом на природном језику и имплементационих детаља у програмском коду. Очигледно је да се програмски код и природни језик значајно разликују и да су хетерогени у погледу лексичких токена и језичке структуре. Они једино могу бити семантички повезани [9]. Стога, ефикасан претраживач програмског кода треба да разуме семантичко значење упита на природном језику, али и програмског кода.

Семантичка претрага програмског кода користи методе обраде природног језика – ОПЈ (енгл. *Natural Language Processing*) и машинског/дубоког учења како би разумела намеру програмера који пише упит и како би му вратила релевантне делове програмског кода. Ова техника покушава да премости јаз између језика који се користи у коду (програмског језика који се често пише у скраћеницама и високо је технички) и природног језика који је погодан за описивање непрецизираних концепата и идеја [10]. Како би ови модели успешно функционисали потребно их је обучити на одговарајућем задатку помоћу великих скупова података. Најчешће тај задатак подразумева да модели уче семантику програмског кода тако што покушавају да и код и одговарајући текст на природном језику репрезентују сличним векторима у заједничком векторском простору. Тада се процес претраге своди на рачунање сличности између вектора упита и вектора блокова кода [10]. Како не би морали ручно да праве скупове парова <упит на природном језику, блок програмског кода> истраживачи за моделирање ове везе обично користе документационе коментаре због тога што они на природном језику описују код на који се односе. Углавном се из софтверских пројеката аутоматизовано извлаче описи метода и њихов код [9, 10, 11, 12, 13]. Поред коментара, у неким истраживањима модели уче и помоћу структурних информација из програмског кода, као што су апстрактно синтаксно стабло или граф контроле тока [14, 15, 16].

Предмет истраживања предложене докторске дисертације је развој система за семантичку претрагу програмског кода. Овај систем ће, као и многи постојећи системи, користити коментаре из програмског кода за моделирање везе између програмског кода и природног језика. Да би одабрали погодне коментаре предмет истраживања ће бити и технике класификације коментара из програмског кода. Фокус ће бити на софтверским пројектима

писаним са коментарима писаним на српском језику и са кодом писаним на различитим програмским језицима.

Циљ рада је реализација система за семантичку претрагу програмског кода помоћу техника дубоког учења и обраде природних језика. У оквиру тог система ће бити реализован и класификатор коментара из програмског кода. Систем ће бити прилагођен за рад са софтверским пројектима писаним на српском језику и различитим програмским језицима.

Истраживање ће имати значајан утицај на широк спектар заинтересованих страна у развоју софтвера на српском језику с обзиром на то да су постојећа решења фокусирана на енглески језик. Нови скупови података ће додатно обогатити тренутно ограничене ресурсе за ОПЈ на српском језику.

- [1] Seymour, Tom, Dean Frantsvog, Satheesh Kumar. 2011. „History of search engines.“ *International Journal of Management & Information Systems (IJMIS)* 15 (4): 47-58. doi:10.19030/ijmis.v15i4.5799.
- [2] Carroll, Jordan. 2022. *The global developer community | The State of the Octoverse*. Последњи приступ Април 24, 2023. <https://octoverse.github.com/2022/developer-community>.
- [3] Inoue, Katsuro, Yuya Miyamoto, Daniel M. German, Takashi Ishio. 2020. „Code Clone Matching: A Practical and Effective Approach to Find Code Snippets.“ *arXiv*. doi:10.48550/arXiv.2003.05615.
- [4] Di Grazia, Luca, Michael Pradel. 2023. „Code Search: A Survey of Techniques for Finding Code.“ *ACM Computing Surveys (Association for Computing Machinery)* 55 (11). doi:10.1145/3565971.
- [5] Sadowski, Caitlin, Kathryn T. Stolee, Sebastian Elbaum. 2015. „How Developers Search for Code: A Case Study.“ *Proceedings of the 2015 10th Joint Meeting on Foundations of Software Engineering*. Association for Computing Machinery. 191-201. doi:10.1145/2786805.2786855.
- [6] Linstead, Erik, Sushil Bajracharya, Trung Ngo, Paul Rigor, Cristina Lopes, Pierre Baldi. 2009. „Sourcerer: mining and searching internet-scale software repositories.“ *Data Mining and Knowledge Discovery (Springer)* 18: 300-336. doi:10.1007/s10618-008-0118-x.
- [7] McMillan, Collin, Mark Grechanik, Denys Poshyvanyk, Qing Xie, Chen Fu. 2011. „Portfolio: Finding Relevant Functions and Their Usage.“ *Proceedings of the 33rd International Conference on Software Engineering*. Association for Computing Machinery. 111-120. doi:10.1145/1985793.1985809.
- [8] Lv, Fei, Hongyu Zhang, Jian-guang, Wang, Shaowei Lou, Dongmei Zhang, Jianjun Zhao. 2015. „CodeHow: Effective Code Search Based on API Understanding and Extended Boolean Model (E).“ *Proceedings of the 30th IEEE/ACM International Conference on Automated Software Engineering*. 260-270. doi:10.1109/ASE.2015.42.
- [9] Gu, Xiaodong, Hongyu Zhang, Sunghun Kim. 2018. „Deep Code Search.“ *Proceedings of the 40th International Conference on Software Engineering*. Association for Computing Machinery. 933-944. doi:10.1145/3180155.3180167.
- [10] Husain, Hamel, Ho-Hsiang Wu, Tiferet Gazit, Miltiadis Allamanis, Marc Brockschmidt. 2019. „Codesearchnet challenge: Evaluating the state of semantic code search.“ *arXiv*. doi:10.48550/arXiv.1909.09436.
- [11] Hu, Haize, Jianxun Liu, Xiangping Zhang, Ben Cao, Siqiang Cheng, Teng Long. 2023. „A mutual embedded self-attention network model for code search.“ *Journal of Systems and Software (Elsevier)* 198: 111591. doi:10.1016/j.jss.2022.111591.
- [12] Wu, Chen, Ming Yan. 2022. „Learning Deep Semantic Model for Code Search using CodeSearchNet Corpus.“ *arXiv*. doi:10.48550/arXiv.2201.11313.
- [13] Fang, Sen, You-Shuai Tan, Tao Zhang, Yeping Liu. 2021. „Self-attention networks for code search.“ *Information and Software Technology (Elsevier)* 134: 106542. doi:10.1016/j.infsof.2021.106542.
- [14] Wan, Yao, Jingdong Shu, Yulei Sui, Guandong Xu, Zhou Zhao, Jian Wu, Philip Yu. 2019. „Multi-modal attention network learning for semantic source code retrieval.“ *Proceedings of the 34th IEEE/ACM International Conference on Automated Software Engineering*. 13-25. doi:10.1109/ASE.2019.00012.
- [15] Gu, Jian, Zimin Chen, Martin Monperrus. 2021. „Multimodal Representation for Neural Code Search.“ *Proceedings of the 2021 IEEE International Conference on Software Maintenance and Evolution*. 483-494. doi:10.1109/ICSME52107.2021.00049.
- [16] Yu, Dongjin, Quanxin Yang, Xin Chen, Jie Chen, Yihang Xu. 2023. „Graph-based code semantics learning for efficient semantic code clone detection.“ *Information and Software Technology (Elsevier)* 156: 107130. doi:10.1016/j.infsof.2022.107130.

3. Полазне хипотезе

- Коментари из програмског кода описују семантику кода на који се односе и могу се користити као помоћ при развоју система за семантичку претрагу кода.
- Нису сви коментари из програмског кода једнако значајни за задатак семантичке претраге кода.
- Могуће је конструисати репрезентацију програмског кода која у себи садржи семантичко значење.

- Програмски код се може семантички претраживати помоћу техника обраде природног језика.
- Могуће је реализовати систем за семантичку претрагу кода који је писан на различитим програмским језицима.
- Могуће је реализовати систем за семантичку претрагу кода из софтверских пројеката писаних на српском језику који ће имати упоредиве перформансе са већ постојећим системима који се баве пројектима на енглеским језиком.

4. Научне методе истраживања

У предложеном истраживању ће бити развијен систем за семантичку претрагу програмског кода. Овај систем ће бити прилагођен за рад са софтверским пројектима на српском језику и различитим програмским језицима (C, C++, C#, Java, JavaScript/TypeScript, Python, PHP и SQL). У основи система ће се налазити модели дубоког учења који треба да науче семантику програмског кода.

Као и у постојећим истраживањима, за обучавање ових модела биће коришћени парови <документациони коментар, блок програмског кода>. Међутим, нису сви коментари у коду коментари који описују семантику (функционалност) кода. На пример, коментар у оквиру кога се наводи име аутора кода или лиценца нису значајни за семантичку претрагу. Због тога ће први део истраживања обухватити подсистем за класификацију коментара из програмског кода. Овај подсистем ће аутоматски класификовати коментаре у две категорије: *функционалне* коментаре и *нефункционалне* коментаре. Систем ће бити базиран на принципима надгледаног машинског учења и користиће напредне трансформер моделе за класификацију. Због тога што тренутно не постоји јавно доступан скуп коментара на српском језику који су анотирани на потребан начин, такав скуп ће бити креиран у оквиру истраживања. Коментари ће бити прикупљени из софтверских пројеката писаних на српском језику и на свим наведеним програмским језицима. Обучени модели ће бити детаљно евалуирани.

Након класификације коментара из програмског кода фокус ће бити на дизајнирању система за семантичку претрагу кода. И овај систем ће бити заснован на трансформер моделима. Овакви модели захтевају велике количине података за обучавање. Због тога ће бити креиран додатни скуп података који ће садржати парове <коментар, блок програмског кода>. Ови парови ће бити аутоматски извучени из софтверских пројеката отвореног кода писаних на српском језику и различитим програмским језицима и филтрирани помоћу реализованог класификатора. Систем за семантичку претрагу ће, како и слични системи из савремених истраживања, учити семантику програмског кода тако што ће учити репрезентације кода и текста на природном језику у заједничком векторском простору. Циљ је да репрезентације блока кода и њему одговарајућег коментара буду близу у том векторском простору, а да репрезентације семантички различитих блокова кода и текстова буду што даље.

По узору на [10] биће креиран и скуп за евалуацију семантичке претраге кода на српском језику. Биће разматрани српски преводи 99 честих упита који су представљени у поменутом истраживању. Парови <упит, блок програмског кода> биће ручно анотирани оценом степена сличности на скали 0-3 где 0 означава потпуну ирелевантност а 3 потпуно подударање између упита и блока програмског кода. На крају ће реализовани систем за семантичку претрагу бити опсежно евалуиран помоћу овог новог скупа података.

5. Очекивани научни допринос

- Систематски преглед литературе и метода из области класификације коментара програмског кода и семантичке претраге програмског кода.

- Креирање нових скупова података за обучавање модела дубоког учења приликом решавања проблема класификације коментара из програмског кода и семантичке претраге програмског кода.
- Развој аутоматизованог подсистема за класификацију коментара из програмског кода на српском језику и на различитим програмским језицима, његова евалуације и поређење са резултатима релевантних истраживања.
- Развој система за семантичку претрагу програмског кода прилагођеног за рад са софтверским пројектима на српском језику и различитим програмским језицима, његова евалуације и поређење са резултатима релевантних истраживања.
- Нови скуп података за евалуацију система за семантичку претрагу програмског кода на српском језику.

6. План истраживања и структура рада

Истраживање ће бити спроведено кроз неколико фаза:

1. Преглед литературе – Прва фаза у истраживању обухвата упознавање и анализу постојећих решења за класификацију коментара из програмског кода и семантичку претрагу програмског кода. Фокус ће бити на истраживањима која користе машинско и дубоко учење.
2. Развој подсистема за класификацију коментара – Како нису сви коментари релевантни за задатак семантичке претраге програмског кода у овој фази ће бити реализован класификатор коментара заснован на напредним моделима машинског учења. Ова фаза обухвата дизајнирање нове таксономије за класификацију коментара, креирање скупа података за обучавање модела, развој модела за класификацију и њихову евалуацију.
3. Развој система за семантичку претрагу кода – Трећа фаза обухвата дизајн и имплементацију система за семантичку претрагу кода заснованог на дубоком учењу. За потребе обучавања ће бити креиран нови скуп података, с обзиром да тренутно не постоје ресурси за српски језик.
4. Евалуацију система за семантичку претрагу кода – У овој фази ће бити креиран и ручно аотиран нови скуп података који ће се користити за евалуацију система за семантичку претрагу програмског кода на српском језику. Реализовани систем за семантичку претрагу кода ће бити опсежно евалуиран на овом скупу података.
5. Дисеминација и верификација добијених резултата кроз научне публикације у међународним часописима и на међународним конференцијама.

7. Закључак и предлог

На основу приложене документације и непосредног увида у досадашњи рад кандидаткиње Марије Костић, мастер инжењера електротехнике и рачунарства и информационих технологија, Комисија закључује следеће:

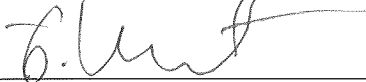
1. Кандидаткиња испуњава све формалне и суштинске услове прописане Законом о високом образовању, Статутом Универзитета у Београду и Статутом Електротехничког факултета у Београду за израду докторске дисертације.
2. Предложена тема **„Интелигентни систем за семантичку претрагу програмског кода“** јесте научно заснована, а резултати који ће проистећи изградом ове докторске дисертације ће представљати значајан и оригиналан допринос међународној научној заједници. Тема припада ужој научној области „Вештачка интелигенција“.

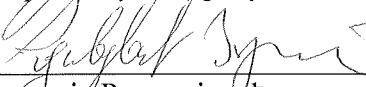
На основу претходно изложеног, Комисија са задовољством предлаже Наставно-научном већу Електротехничког факултета у Београду да кандидаткињи Марији Костић одобри израду докторске дисертације **„Интелигентни систем за семантичку претрагу програмског кода“** (енгл. *„Intelligent system for semantic search of source code“*), под

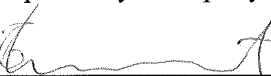
менторством др Бошка Николића, редовног професора Електротехничког факултета у Београду.

У Београду, 09.06.2023.

ЧЛАНОВИ КОМИСИЈЕ:

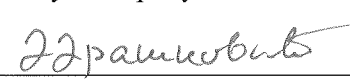

др Бошко Николић, редовни професор
Универзитет у Београду – Електротехнички факултет


др Захарије Радивојевић, ванредни професор
Универзитет у Београду – Електротехнички факултет


др Борис Делибашић, редовни професор
Универзитет у Београду – Факултет организационих наука


др Татјана Лутовац, редовни професор
Универзитет у Београду – Електротехнички факултет


др Вук Батановић, научни сарадник
Иновациони центар Електротехничког факултета
Универзитета у Београду


др Дражен Драшковић, доцент
Универзитет у Београду – Електротехнички факултет