Република Србија

КРИМИНАЛИСТИЧКО-ПОЛИЦИЈСКИ УНИВЕРЗИТЕТ

Број: 43/19

Датум: 2. 4. 2025. год

ВЕЋУ ДЕПАРТМАНА ИНФОРМАТИКЕ И РАЧУНАРСТВА КРИМИНАЛИСТИЧКО-ПОЛИЦИЈСКОГ УНИВЕРЗИТЕТА У БЕОГРАДУ

Предмет: Извештај о урађеној докторској дисертацији кандидата Славише Ж. Илића

Поштовани чланови Већа департмана информатике и рачунарства,

Одлуком Већа научних области природно-математичких и техничко-технолошких студија Криминалистичко-полицијског универзитета 16 бр. 85/2-3-2025 од 19.03.2025. одређени смо за чланове Комисије за оцену докторске дисертације докторанда **Славише Ж. Илића** под насловом **„Унапређење динамичке анализе злонамерног софтвера засновано на проширеном скупу обележја и машинском учењу“.**

После прегледа достављене дисертације, Комисија подноси следећи

ИЗВЕШТАЈ

1. УВОД

1.1 Хронологија одобравања и израде дисертације

Кандидат је уписао докторске академске студије информатике 2020/21. на Департману информатике и рачунарства Криминалистичко-полицијског универзитета. Након што је положио све испите предвиђене студијским програмом са просечном оценом 9,86, поднео је пријаву теме докторске дисертације бр. 43/19 од 07.10.2024. године, под називом „Унапређење динамичке анализе злонамерног софтвера засновано на проширеном скупу обележја и машинском учењу“ у научном пољу Природно-математичке науке, научној области Рачунарске науке и ужој научној области Рачунарство.

Веће научних области природно-математичких и техничко-технолошких студија Криминалистичко-полицијског универзитета прихватило је предложену тему докторске

дисертације и за ментора именовало проф. др Милана Гњатовића (одлука 16 бр. 85/6-3-2024 од 20.12.2024.).

У складу с чл. 26 ст. 1 Правилника о докторским студијама, 22 бр. 79/10-4-2022 од 19.09.2022., кандидат Славиша Илић је 05.03.2025. поднео завршену докторску дисертацију на оцену, уз прилагање потребних доказа (бр. 43/9).

У складу с чл. 26 ст. 2 Правилника о докторским студијама, ментор је дао писану сагласност бр. 43/10 од 05.03.2025., којом је констатовао да је докторска дисертација коју је кандидат поднео подобна за оцену и да кандидат има објављен најмање један рад у научном часопису са импакт-фактором на СЦИ листи, односно СЦИе листи, који је повезан са садржајем докторске дисертације и у којем је кандидат први аутор.

У складу с чл. 28, ст. 1 Правилника о докторским студијама, извршена је провера оригиналности достављене докторске дисертације. У складу с чл. 28, ст. 2 Правилника о докторским студијама, ментор је 07.03.2025. руководиоцу надлежног департмана доставио позитивно писано мишљење о процени оригиналности докторске дисертације.

У складу с чл. 29, ст 2 Правилника о докторским студијама, а на основу предлога руководиоца надлежног департмана бр 43/12 од 12.03.2025., Веће научних области природно-математичких и техничко-технолошких студија Криминалистичко-полицијског универзитета је одлуком 16 бр. 85/2-3-2025 од 19.03.2025. образовало Комисију за оцену докторске дисертације докторанда Славише Илића у следећем саставу:

- др Петар Чисар, председник, редовни професор Криминалистичко-полицијског универзитета,
- др Кристијан Кук, члан, редовни професор Криминалистичко-полицијског универзитета,
- др Иван Тот, члан, ванредни професор Војне Академије Универзитета одбране.

1.2 Научна област дисертације

Докторске студије информатике на Криминалистичко-полицијском универзитету акредитоване су у оквиру образовно-научног поља Природно-математичких наука, научна област Рачунарске науке, у складу с чим Комисија констатује да је предметна докторска дисертација пријављена у научној области за коју је Криминалистичко-полицијски универзитет матичан.

Предложени ментор др Милан Гњатовић јесте редовни професор на Департману информатике и рачунарства Криминалистичко-полицијског универзитета, изабран у ужој научној области Рачунарство. Задовољава све услове за ментора и налази се на Листи ментора Департмана информатике и рачунарства Криминалистичко-полицијског универзитета. Коаутор је 18 радова објављених у научним часописима са импакт-фактором на СЦИ/СЦИе листи. Увидом у базу SCOPUS установљено је 384 хетероцитата и остварени h-индекс 12.

1.3 Биографски подаци о кандидату

Славиша Ж. Илић рођен је 1982. године у Чукојевцу, у околини Краљева. Завршио је Војну гимназију 2001. године са одличним успехом, и петогодишње студије на Војној академији, смер Информатика, 2006. године са просеком 8,38. Докторске академске студије информатике на Департману информатике и рачунарства Криминалистичко-полицијског универзитета уписао је 2020. године.

Радио је на различитим позицијама у Војсци Србије и Министарству одбране као дипломирани инжењер информатике и асистент на Катедри информатике на Војној академији.

2 ОПИС ДИСЕРТАЦИЈЕ

2.1 Садржај дисертације

Докторска дисертација је изложена у седам поглавља и има следећу структуру:

1. „Увод“,
2. „Преглед релевантних истраживања“,
3. „Нови корпус за динамичку анализу злонамерног софтвера“,
4. „Класификовање злонамерног софтвера“,
5. „Анализа обележја“,
6. „Дискусија“,
7. „Закључак“.

Дисертација је изложена на 127 страна формата А4, и садржи 16 слика, 17 табела и 59 литературних навода. После насловне стране стоје превод насловне стране на енглески језик, страна с подацима о ментору, захвалница, сажетак на српском и енглеском језику, а потом следе садржај, попис изабраних скраћеница, списак табела, списак слика, горенаведена поглавља, списак референци, биографија кандидата, изјава о ауторству, изјава о истоветности штампане и електронске верзије докторске дисертације и изјава о коришћењу.

2.2 Кратак приказ појединачних поглавља

Наслов докторске дисертације је јасно формулисан и сажето дефинише тематику и садржај дисертације.

У уводном поглављу представљени су проблем и предмет научног истраживања и дат је хипотетички оквир истраживања. Истакнуто је да злонамерни софтвери представљају изазов у области информационе безбедности, јер узрокују материјалне трошкове, губитак информација и штету по углед жртве. Као проблем научног истраживања, ова дисертација посматра специфични аспект аутоматског интерпретирања извештаја о динамичкој анализи софтвера, који се односи на унапређење аутоматске динамичке анализе злонамерног софтвера у контексту машинског учења. Међутим, релевантни приступи динамичкој анализи софтвера засновани на машинском учењу доминантно су фокусирани на обележја која се односе на АПИ-позиве. Два главна недостатка приступа фокусираних искључиво на обележја која се односе на АПИ-позиве јесу следећа. Динамичка обележја која не припадају скупу АПИ-позива, попут обележја која рефлектују приступање фајловима на диску, промене у систему регистара оперативног система, мрежни саобраћај итд., јесу занемарена, иако потенцијално могу да буду релевантна за детектовање злонамерног софтвера. Значајан број злонамерних и безопасних софтвера не генерише АПИ-позиве, што упућује на извођење додатних динамичких обележја. Предмет научног истраживања у овој дисертацији, односи се на проширење скупа динамичких обележја, у циљу унапређења перформанси аутоматског детектовања злонамерног софтвера у контексту машинског учења. Општа хипотеза овог истраживања јесте да постоје динамичка обележја софтвера која су релевантна за аутоматско детектовање злонамерних софтвера, а не припадају скупу АПИ-позива. Посебна хипотеза овог истраживања јесте да проширени скуп динамичких обележја софтвера унапређује тачност система за аутоматско детектовање злонамерних софтвера у односу на систем који би био заснован само на скупу обележја која се односе на АПИ-позиве. На крају, прво поглавље даје кратке прегледе доприноса и структуре дисертације.

У другом поглављу дат је приказ релевантних истраживања у области динамичке анализе софтвера. Фокус разматрања примарно је стављен на примењене корпусе динамичких обележја и моделе машинског учења, уз систематичан преглед перформанси размотрених модела. Истакнута су два главна смера истраживања која се односе на издвајање обележја. Први смер, који је доминантно заступљен, заснован је на анализи обележја која се односе на АПИ-позиве. Други смер, који је знатно мање заступљен, полази од претпоставке да скуп релевантних обележја превазилази скуп АПИ-позива. Према најбољем сазнању аутора, у овом тренутку су доступне само две публикације у којима је описан овај смер истраживања. Међутим, корпус динамичких обележја у једној од тих публикација није јавно доступан, док корпус из другог истраживања не укључује безопасне узорке и садржи злонамерне узорке из само шест породица злонамерног софтвера. Тиме је аутор успешно указао на потребу за генерисањем новог скупа података, јавно доступног за потребе истраживања, који садржи извештаје о динамичкој анализи безопасни узорака софтвера и злонамерних узорака софтвера из обимнијег скупа породица.

У трећем поглављу описан је поступак генерисања новог скупа података, насталог динамичком анализом понашања реалних злонамерних и безопасних фајлова. Прво је изведена упоредна анализа два окружења за динамичку анализу софтвера – Куку и Драквуф – на основу које је окружење Куку процењено као прикладније за потребе предметног истраживања. Потом је детаљно описано подешавање лабораторијског окружења за динамичку анализу, укључујући избор оперативног система виртуелне машине за анализу, подешавање техника против избегавања извршења и подешавање окружења Пајтон.

Колекција од приближно 430000 узорака злонамерног софтвера преузета је из базе портала Вирус тотал, који се у професионалној заједници сматра референтним извором. Из ове колекције фајлова, без субјективне пристрасности, програмски је одабрано 10465 узорака који су укључени у корпус. Скуп безопасних фајлова добијен је из архиве података о пословној кореспонденцији, инжењерским нацртима и документацији једног активног привредног друштва регистрованог у Републици Србији. Архива се састоји од приближно 74000 безопасних фајлова насталих у реалном окружењу, од којих је програмски и без субјективне пристрасности одабрано 11735 узорака за укључивање у корпус. Новоформирани корпус садржи комплетне и непромењене извештаје које је аутоматски генерисало окружење Куку за поднете безопасне и злонамерне фајлове. Из овог корпуса накнадно је изведен проширени скуп обележја (у даљем тексту: корпус проширених обележја), који по врстама обележја превазилази скуп АПИ-позива. У складу с циљем истраживања да се покаже да је дискриминаторна моћ корпуса проширених обележја већа од дискриминаторне моћи обележја која се односе само на АПИ-позиве, из корпуса проширених обележја додатно је изведен контролни корпус АПИ-позива. Узорци у овом контролном корпусу садрже само обележја која се односе на АПИ-позиве изведене из узорака у корпусу проширених обележја. Корпус проширених обележја садржи 22200 узорака (11735 безопасних и 10465 злонамерних), који се односе на 54 типа фајла, при чему су дате расподеле узорака по типу фајла у корпусу проширених обележја и корпусу АПИ-позива. При томе, 23,78% извештаја не укључује АПИ-позиве, односно 39.5% безопасних фајлова и приближно 6.1% злонамерних фајлова не индукује АПИ-позиве. У апсолутним вредностима: 5280 фајлова не индукује АПИ-позиве, од којих је 4638 безопасних и 642 злонамерна (в. тачку 4.2).

У четвртом поглављу описан је приступ бинарном класификовању извештаја о динамичкој анализи софтвера, тј., сваки узорак се класификује као безопасан или злонамеран. На методолошком нивоу, описани приступ се заснива на ансамблима стабала одлучивања. У експерименту се посматрају следеће независне променљиве:

- корпус података се односи на корпус проширених обележја или на корпус АПИ-позива,
- техника за издвајање обележја се односи на технику ЦВ (енгл. *Count Vector*) или технику ТФ-ИДФ (енгл. *Term Frequency – Inverse Document Frequency*),
- број стабала у ансамблу узима вредности из опсега [1, 100],
- максимална дубина стабла узима вредности из скупа {*None*, 100, 1000}, где *None* означава да дубина стабла није ограничена,
- минималан број узорака потребних да гранање чвора узима вредности из скупа {2, 100, 1000},
- минимални број узорака по листу узима вредности из скупа {1, 100, 1000},
- број обележја који се разматрају приликом гранања чвора узима вредности из скупа{*sqrt*, 0,1, 0,2}, где *sqrt* означава заокружену вредност квадратног корена броја доступних обележја.

Као зависне променљиве, посматрају се: тачност, микропросечне прецизности по класама, макропросечна прецизност, микропросечни одзиви по класама, макропросечни одзив и макропросечна Φ_1-мера. Експеримент је спроведен на следећи начин. Комбинујући вредности прве три независне променљиве (корпуса, технике за издвајање обележја и броја стабала) генерисано је 400 (2x2x100) ансамбала одлучивања. Ови ансамбли су додатно

оптимизовани комбиновањем вредности преостале четири независне променљиве. Резултати експеримента потврђују општу и посебну хипотезу ове дисертације: (i) Постоје динамичка обележја софтвера која су релевантна за аутоматско детектовање злонамерних софтвера, а не припадају скупу АПИ-позива. (ii) Проширени скуп динамичких обележја софтвера унапређује тачност система за аутоматско детектовање злонамерних софтвера у односу на систем који би био заснован само на скупу обележја која се односе на АПИ-позиве (в. тачку 4.2).

У петом поглављу посматра се питање одређивања динамичких обележја која највише доприносе дискриминаторној моћи прототипских система развијених у претходном поглављу. За свако од преко 25 милиона (25066933) обележја из корпуса израчуната је информациона добит. Вредности информационе важности обележја издвојених техникама ТФ-ИДФ припадају опсегу (0, 0,00636), при чему првих 32134 обележја имају вредности информационе добити које нису занемарљиве. Слично, вредности информационе добити за обележја издвојена техником ЦВ припадају опсегу (0, 0,00585), при чему првих 37285 обележја имају вредности информационе добити које нису занемарљиве. Применом прилагођеног алгоритма за адаптивно одређивање прага за бинаризовање дигиталних слика, одређена је вредност прага информационе добити који раздваја најважнија динамичка обележја софтвера. На тај начин издвојено је 175 најважнијих обележја изведених применом технике ЦВ, и 195 обележја изведених применом технике ТФ-ИДФ. Потом је извршено интерпретирање издвојених обележја и њихова расподела у десет категорија, од којих се само једна категорија односи на АПИ-позиве (в. тачку 4.2).

У шестом поглављу дискутована су три главна доприноса спроведеног истраживања: нови корпус података за динамичку анализу софтвера, практично демонстрирање веће дискриминаторне моћи проширеног скупа обележја и анализа издвојених обележја.

Седмо поглавље закључује дисертацију, уз кратки осврт на доприносе дисертације.

Литература садржи 59 прегледно систематизованих библиографских навода.

3. ОЦЕНА ДИСЕРТАЦИЈЕ

3.1 Савременост и оригиналност

Због константног унапређења функционалности злонамерних софтвера, класични технички механизми одбране (попут антивирусних програма, софтверских агента са реактивном компонентом, мрежне баријера и др.), не спречавају, у општем случају, доспеће злонамерних узорака у радно окружење рачунара. У складу с тим, извесна истраживачка пажња већ је посвећена проблему унапређењу аутоматских механизама заштите. Ова дисертација посматра специфични аспект овог проблема, који се односи се на унапређење аутоматске динамичке анализе злонамерног софтвера у контексту машинског учења. Релевантни приступи динамичкој анализи софтвера засновани на машинском учењу доминантно су фокусирани на обележја која се односе на АПИ-позиве. За разлику од њих, приступ описан у овој дисертацији посматра и додатна обележја која се не односе искључиво на АПИ-позиве. У

складу с тим, представљен је нов корпус за динамичку анализу података, са значајно проширеним скупом динамичких обележја. Треба приметити да тренутно не постоје други јавно доступни корпуси комплетних извештаја о динамичкој анализи софтвера са тако великим бројем узорака као корпус представљен у овој дисертацији. Поред тога, експериментално је демонстрирано да проширени скуп динамичких обележја софтвера унапређује тачност система за аутоматско детектовање злонамерних софтвера у односу на систем који би био заснован само на скупу обележја која се односе на АПИ-позиве.

3.2 Осврт на коришћену литературу

Литература садржи 59 прегледно систематизованих библиографских навода, укључујући референце које су кључне за увођење релевантних појмова, концепата, рачунарских алгоритама и модела, и референце које се односе на скорашња научна истраживања у области динамичке анализе софтвера. Литература је адекватно одабрана и одговара тематици ове дисертације.

3.3. Опис и адекватност примењених научних метода

За спровођења научног истраживања у оквиру предметне докторске дисертације примењене су основне и савремене методе за рачунарско моделовање, извођење обележја, надгледано машинско учење, статистичку анализу, упоредну анализу, обраду дигитаних слика, као и експериментални метод. Све методе су адекватно примењене.

3.4 Применљивост остварених резултата

Предложени приступ динамичкој анализи злонамерног софтвера практично је имплементиран и позитивно оцењен на корпусу који садржи 22200 узорака реалних злонамерних и безопасних фајлова.

3.5 Оцена достигнутих способности кандидата за самостални научни рад

У својој докторској дисертацији кандидат је спровео истраживање литературе из предметне области, извршио анализу постојећег стања истраживања и предложио приступ који унапређује динамичку анализу злонамерног софтвера. Тиме је демонстрирао самосталност у истраживачком раду и способност критичког размишљања. Објављивањем резултата у научном часопису са импакт-фактором на СЦИ/СЦИе листи, кандидат је демонстрирао способност за представљање научних доприноса. Комисија је мишљења да је кандидат показао адекватан степен способности за самосталан научноистраживачки рад.

4. ОСТВАРЕНИ НАУЧНИ ДОПРИНОС

4.1 Приказ остварених научних доприноса

Први допринос овог истраживања представља генерисање новог корпуса за динамичку анализу злонамерног софтвера, који садржи извештаје о резултатима динамичке анализе узорака и задовољава следеће карактеристике:

- сви изворни фајлови који су предмет динамичке анализе представљају примере из реалног живота,
- изворни скуп фајлова садржи педесет четири врсте фајлова (тј., различите екстензије),
- корпус садржи 22200 узорака генерисаних спровођењем аутоматске динамичке анализе изворних фајлова у окружењу Куку, са следећом расподелом: 10465 узорака представља резултате динамичке анализе злонамерних софтвера, 11735 узорака представља резултате динамичке анализе безопасних софтвера,
- величина корпуса износи 969GB, при чему злонамерни узорци заузимају 935GB, а безопасни 34GB,
- из корпуса је изведено више од 25 милиона динамичких обележја софтвера и
- корпус је јавно доступан за потребе истраживања.

Иако су злонамерни и безопасни узорци уравнотежено заступљени, величина злонамерних узорака је знатно већа од величине безопасних узорака. Злонамерни узорци су генерисали више активности у окружењу за динамичку анализу од безопасних узорака, па су резултујући извештаји о њиховој динамичкој анализи већи (чак 228 злонамерних софтвера индукује извештаје веће од 1GB).

Други допринос овог истраживања односи се на:

- развој прототипског система за аутоматску динамичку анализу софтвера у контексту бинарног класификовања софтвера на злонамерни и безопасни, заснованог на динамичким обележјима из датог корпуса и ансамблу стабала одлучивања,
- експериментално демонстрирање да проширени скуп динамичких обележја унапређује перформансе оваквог система у односу на систем који би био заснован само на скупу обележја која се односе на АПИ-позиве.

Из корпуса проширених обележја изведен је означени корпус АПИ-позива, који садржи само узорке који индукују АПИ-позиве. За експерименталну потврду веће дискриминаторне моћи проширеног скупа обележја у односу на обележја која се односе на АПИ-позиве, примењен је ансамбл стабала одлучивања. Експериментални резултати потврђују да модели развијени на проширеном скупу динамичких обележја остварују већу тачност (и макропросечну Φ_1-меру) од модела развијених на скупу обележја која се односе на АПИ-позиве.

Трећи допринос овог истраживања односи се на анализу изведених обележја и одређивање типова обележја који су најзначајнији за откривање злонамерног софтвера.

4.2 Критичка анализа резултата истраживања

Предложено унапређење динамичке анализе злонамерних софтвера систематски је и прецизно изложено, концептуално оправдано и адекватно валидирано. Дискусија резултата је адекватна и потврђује валидност представљеног приступа. Закључци донети на бази изложених резултата потврђују значај предложеног приступа динамичкој анализи софтвера.

Број узорака у корпусу генерисаном у оквиру истраживања описаног у овој дисертацији, његова величина и број изведених обележја знатно превазилазе корпусе сличне намене сачињене само од секвенци АПИ-позива. Треба приметити да тренутно не постоје други јавно доступни корпуси са оваквим карактеристикама који би били прикладни за анализу бинарног класификовања злонамерног софтвера засновану на скупу динамичких обележја која, по свом типу, превазилазе АПИ-позиве.

Тачност остварена на проширеном скупу обележја износи 99,74%, а на скупу обележја која се односе на АПИ-позиве 95,56%. Треба нагласити да су у корпус АПИ-позива укључени само узорци који су изведени из фајлова који индукују АПИ-позиве. Другим речима, 23,78% узорака који не укључују АПИ-позиве (тј., 39,5% безопасних фајлова и приближно 6,1% злонамерних фајлова) није укључено у овај корпус. Да су и ови узорци били укључени у анализу, тачност анализе засноване на АПИ-позивима била би смањена, а предност примене проширеног скупа обележја још наглашенија.

Експертским класификовањем најважнијих динамичких обележја утврђено је да заступљеност обележја која се односе на АПИ-позиве не прелази 32,39% свих обележја издвојених техником ЦВ, односно 33.37% процената свих обележја издвојених техником ТФ-ИДФ. Ово упућује на закључак да је проширење скупа динамичких обележја значајно. Осим тога, предност примене технике ТФ-ИДФ у односу на технику ЦВ јесте што она резултује не само већим бројем изведених обележја, већ и већом разноврсношћу њихових типова.

4.3 Верификација научних доприноса

Кандидат Славиша Илић је први аутор на раду објављеном у истакнутом међународном часопису (М22):

- Ilić S, Gnjatović M, Tot I, Jovanović B, Maček N, Gavrilović Božović M. Going beyond API Calls in Dynamic Malware Analysis: A Novel Dataset. Electronics. 2024; 13(17):3553. https://doi.org/10.3390/electronics13173553.

Овај рад припада научној области докторске дисертације и тематски одговара проблему и предмету научног истраживања дефинисаним у предлогу теме докторске дисертације. За одређивање М-категорије часописа коришћен је JCR Science Edition, за период од две године пре публиковања и година публиковања, и то за ону годину у којој је часопис најбоље рангиран, у складу са Прилогом 2 Правилника о поступку и начину вредновања, и квантитативном исказивању научноистраживачких резултата истраживача ("Сл. Гласник РС", бр. 24/2016 и 21/2017).

5. ЗАКЉУЧАК И ПРЕДЛОГ

Комисија за оцену докторске дисертације „Унапређење динамичке анализе злонамерног софтвера засновано на проширеном скупу обележја и машинском учењу“ кандидата г. Славише Ж. Илића сматра да је предметна дисертација адекватно методолошки постављена, правилно спроведена, заснована на актуелним сазнањима, урађена у складу са образложењем наведеним у пријави теме и да садржи све елементе специфициране Правилником о докторским студијама Криминалистичко-полицијског универзитета у Београду.

На основу свега изложеног, имајући у виду научне резултате истраживања, практичну примену истраживања, методолошки оквир и научни и друштвени допринос, Комисија даје позитивну оцену предметне докторске дисертације и Већу департмана информатике и рачунарства Криминалистичко-полицијског универзитета у Београду

ПРЕДЛАЖЕ

да се докторска дисертација „Унапређење динамичке анализе злонамерног софтвера засновано на проширеном скупу обележја и машинском учењу“ кандидата г. Славише Ж. Илића прихвати, изложи на увид јавности и упути на коначно усвајање Већу научних области природно-математичких и техничко-технолошких студија Криминалистичко-полицијског универзитета.

У Београду,

ЧЛАНОВИ КОМИСИЈЕ

др Петар Чисар, председник, редовни професор,
Криминалистичко-полицијски универзитет у Београду

др Кристијан Кук, члан, редовни професор
Криминалистичко-полицијски универзитет у Београду

др Иван Тот, члан, ванредни професор, Војна академија,
Универзитет одбране