

Рецензия

по процедура за защита на дисертационен труд на тема:

„Интелигентни информационни системи в биоинформатиката: семантично интегриране, анализ и класификация на биомедицински данни“

за придобиване на

образователна и научна степен „доктор“

от

кандидат: **Илиян Недков Михайлов**,

Област на висше образование: **4. Природни науки, математика и информатика**

Професионално направление: **4.6. Информатика и компютърни науки**

Докторска програма: **„Информационни технологии – Био и медицинска информатика“**, катедра: **„Информационни технологии“**, Факултет по математика и информатика (ФМИ), Софийски университет „Св. Климент Охридски“ (СУ).

Научен ръководител: **доц. д-р Димитър Иванов Василев**

Рецензията е изготвена от: **доц. д-р Светла Николова Бойчева**, Институт по информационни и комуникационни технологии, Българска академия на науките в качеството ми на член на научното жури, съгласно Заповед № РД 38-255 /02.06.2021 г. на Ректора на Софийския университет.

1. Обща характеристика на дисертационния труд и представените материали

Дисертационният труд е с обем 186 страници, в т.ч. 165 стр. основно тяло и 21 страници съдържание, списък на фигури и таблици, речник на термините, използвана литература, заглавни страници. Текстът е организиран в увод (глава 1), 4 глави и заключение, като трябва да се отбележи, че 5-та глава съдържа авторската справка. Използваната литература се състои от 151 източника на английски език. В основната си част използваните източници са публикувани през последните 10 години. Използваната литература е цитирана по подходящ начин в текста на дисертацията, но трябва да се отбележи, че има известни пропуски. Например, за източници [21] и [22] липсва позовавания в текста на дисертационния труд.

2. Данни и лични впечатления за кандидата

Докторантът Илиян Михайлов е завършил през 2015г. ОКС „Бакалавър“, специалност Компютърни системи и технологии, професионална квалификация „Компютърен инженер“ в Технически университет-Габрово. През 2017 г. е завършил ОКС „Магистър“, специалност Компютърни системи и технологии, професионална квалификация „Магистър -Компютърен инженер“ в Технически университет-Габрово. От м. март 2014 до м. април 2017 е работил като софтуерен разработчик в Vulpro. В периода 2018-2021 г. е бил докторант във Факултета по математика и информатика на Софийски университет „св. Климент Охридски“, докторска програма: „Информационни технологии – Био и медицинска информатика“, към катедра „Информационни технологии“. Паралелно с това от м. април 2017 г до момента работи и като старши софтуерен разработчик в софтуерна компания SAP Labs България.

Познавам лично кандидата от негови участия в научни форуми и имам отлични впечатления от него като качества на отговорен човек и добре изграден специалист в областта на компютърните науки, който умее да извършва научна дейност и да представя научни резултати с високо качество.

3. Съдържателен анализ на научните и научно-приложните постижения на кандидата, съдържащи се в представения дисертационен труд и публикациите към него, включени по процедурата

Целта на дисертационния труд е създаването на методология и практическата ѝ реализация за интелигентно интегриране на биомедицински данни и техния анализ, използвайки средства на информатиката, информационните технологии, биоинформатиката и изкуствения интелект.

В **увода (първа глава)** е представена мотивацията и значимостта на решаваният научен проблем. Дефинирани са целта на дисертационния труд и произтичащите от нея задачи, както и е описана структурата на изложението.

Във **втора глава** е направен обзор, представена е теоретичната основа и постановка на задачите.

В **трета глава** са описани модели за съхранение и интегриране на данни. Представени са методите за хоризонтално и вертикално интегриране на хетерогенни данни. Представена е методология за семантично интегриране на биомедицински данни от различни заболявания. Описано е приложението на свързани данни на ба-

зата на онтологии при създаване на съветваща система в областта на здравеопазването. Описан е нов модел за прогнозиране на преживяемост на пациенти болни от рак. Представено е приложение на методи за машинно самообучение при оценка точността на предсказване на протеинови структури. Представен е модел за компресиране на омикс данни. Представено е приложение на Gradient Boosting Machine, Random Forest и невронни мрежи при решаване на задачата за предсказване на антимикробна резистентност в метагеномни данни.

В **четвърта глава** е дадено описание на софтуерният модул за интегриране на хетерогенни данни. Описани са и разработените модели за предсказване, компресия и класификация на различни видове биомедицински данни.

В **пета глава** е представена авторската справка.

В **заключението** е направено обобщение на представените научни резултати и зададени насоки за бъдеща работа.

В авторската справка са описани 5 научни и 6 приложни приноса. Добре би било част от научните приноси да се преформулират, прецизират и доуточнят, защото по принцип звучат много общо и биха могли да доведат до заблуда, че са предложени решения на проблемите в общия случай, а не в конкретен прекалено частен случай, за който е направено описание в дисертационния труд.

4. Аprobация на резултатите

От формална гледна точка, съгласно Правилника за прилагане на ЗРАСРБ, Постановление № 26 от 13.02.2019 г. за изменение и допълнение на ППЗРАСРБ, касаещо Област 4. Природни науки, математика и информатика, проф. направления 4.1., 4.2., 4.3., 4.4., 4.5., 4.6., и Правилника за специфичните условия и реда за придобиване на научни степени и заемане на академични длъжности в СУ „Св. Климент Охридски” се изисква докторантът да има минимум 30 точки от група Г (показатели 5-10).

Докторантът Илиян Михайлов е представил общо 1 патент (С1) и 9 публикации (С2-С10), като всички публикации са на английски език в рецензирани издания:

- Публикация С9 е в сборник от абстракти и е изключена от разглежданията.
- Публикациите в списания са следните:
 - С2 (Biology Direct, Scopus-Q1, SJR 1.51; Web of Science-Q1, IF 4.54)
 - С4 (Information, MDPI, Scopus-Q3, SJR 0.353; Web of Science без кuartил и без IF)
 - С8 (Frontiers in Genetics, Scopus-Q2, SJR 1.413; Web of Science-Q2, IF 4.599)

- Публикациите в сборник трудове от научни конференции са както следва:
 - C3 (Lecture Notes in Computer Science – Springer, Scopus-Q2, SJR 0.283; Web of Science-Q4, без IF)
 - C5 (Lecture Notes in Computer Science – Springer, Scopus- Q3, SJR 0.249; Web of Science-Q4, без IF)
 - C7 (Lecture Notes in Computer Science – Springer, Scopus-Q2, SJR 0.427; Web of Science-Q4, без IF)
 - C6 (IEEE Explore, Scopus - без SJR)
 - C10 (CEUR-WS, Scopus- без квартил, SJR 0.177)

Всички представени публикации са в съавторство, както с научният ръководител на докторанта, така и с други български и чуждестранни учени. На четири от публикациите Илиян Михайлов е първи автор (C2, C3, C4, C7). Докторантът не е представил нито една самостоятелна публикация. Научните резултати са представени в 11 доклада на научни форуми.

Съгласно Постановление № 26 от 13.02.2019 г. за изменение и допълнение на ППЗРАСРБ се зачитат само научни публикации в издания, които са реферирани и индексирани в световноизвестни бази данни с научна информация (Web of Science и Scopus, Zentralblatt, MathSciNet, ACM Digital Library, IEEE Xplore и AIS eLibrary)

Така за точка Г (показатели 5 до 10) се отчитат само 8 от публикациите (C2-C8, C10) на Илиян Михайлов. Тогава сумарният брой точки по този показател покрива и многократно надвишава минималното изискване от 30 т.

Забелязани са 33 цитирания в Scopus на публикации на докторанта, представени в дисертационния труд, като са изключени автоцитиранията. Публикациите на докторанта в реномирани международни научни издания, както и множеството им цитирания от световни учени, показват значимостта на получените резултати от научните изследвания за областта.

Представените научните трудове по тази процедура, отговарят на минималните национални изисквания (по чл. 2б, ал. 2 и 3 на ЗРАСРБ) и съответно на допълнителните изисквания на СУ „Св. Климент Охридски“ за придобиване на образователна и научна степен „доктор“ в научната област и професионално направление на процедурата.

Представените резултати представят оригинален научен труд и не повтарят съществуващи изследвания представяни в други процедури за придобиване на научни степени и заемане на академични длъжности.

Предоставени са данни за проверка за плагиатство и няма доказано по законоустановения ред плагиатство в представения дисертационен труд и научни трудове по тази процедура.

Според мен приносет на кандидата при колективните публикации е ясен и съществен. Той личи от засегнатата тематика. На четири от публикациите докторанта е първи автор.

5. Качества на автореферата

Беше ми предоставен автореферат на български и английски език. Той съдържа 37 страници и отразяват основните глави и резултати от дисертацията. Прави впечатление, че при изписване на името на кандидата в автореферата на английски език има разлика с изписването на името, което е използвано в публикациите. Това би могло да доведе до известно объркване и би било добре да е унифицирано.

6. Критични бележки и препоръки

Бих направила някои технически бележки, които не намаляват стойността на представения дисертационен труд от гледна точка на постигнатите резултати от кандидата, но биха способствали за подобряването на изложението:

1. Според мен целта на дисертационния труд не е много добре дефинирана – нуждае се от известно прецизиране в посока акцент върху научния проблем, който се решава, както и дефиниране на предмет и обект на изследването, дефиниране на работна хипотеза, от която да произтичат задачите на дисертационния труд. Всички дефинирани задачи са прекалено амбициозни и покриват много нерешени проблеми в областта на биомедицинската информатика, но липсва взаимовръзката между тях при толкова общо дефинирана цел. Единственото свързващо звено между различните подзадачи е областта на приложение - биомедицинска информатика. Това е и една от причините получените резултати да не са задълбочени изследвания за един проблем, а да са разпръснати в твърде широк спектър от теми.
2. Текстът на дисертационният труд не е изчистен от синтактични и граматически грешки. Изобилства използването на жаргонни думи и чуждици, типично използвани в информационните технологии. Въпреки добавянето на речник на използваните термини, за съжаление и там част от термините са оставени без превод на български език. Прави впечатление и, че за част от термините е

използван неправилен превод на български език, например за основният термин „граф“ на места е използван термина „графика“, което е доста затрудняващо четенето и разпознаването му в определен контекст. Изчистването на текста от грешки и подобряване на стилът на изложение би способствало за увеличаване на качеството на представеният дисертационен труд.

3. Изложението на дисертационния труд е малко хаотично и се отклонява от стандартния формат – обзор на подходите, описание на предлаганият подход и илюстрация на неговото приложение. В голямата си част изложението акцентира на това какво по принцип би могло да се направи за решаване на проблема, но без достатъчно конкретика как точно се прави това в предлаганото решение. Като цяло се набляга повече на използваните софтуерните компоненти и реализация, отколкото на описание на предложените методи и алгоритмични решения. Липсват и достатъчно конкретни примери за илюстрация на предложените подходи. При методите за машинно самообучение обикновено се представят експериментални резултати с еталони (benchmark dataset), за да може да се направи обективна оценка на подобрението от предложения подход. Трудно се разграничава обзорната част на методите и личният принос на докторанта в изложението на дисертационния труд.
4. Относно предложеният подход за интеграция на хетерогенни данни не са засегнати три основни въпроса: (1) приложение на различни методи за съпоставяне на данни от хетерогенни източници (entity matching); (2) нормализация на данни при автоматичното съпоставяне към стандартни класификации и онтологии; (3) методите за извличане на информация – основни обекти и релация между тях при обработка на полуструктурирани данни, които съдържат текстови полета. Първият проблем има в частност отношение и към втория - как да се съпоставят понятия от две бази от данни, които са класифицирани според различни онтологии. Третият проблем предполага използването на подходи от ИИ за обработка на естествен език, като извличане на информация и текстово базирани методи за класификация. Тези проблеми по принцип в общия случай не са решени в общия случай, но тяхното адресиране в дисертационния труд би дало възможност да се даде по-обективна оценка на общността и иновативността на предлаганото решение. Прецизиране на описанието на приносите на докторанта с ясно пояснение за ограниченията на

предлаганият подход за интеграция на хетерогенни данни би било от полза за ясно идентифициране на потенциални негови приложения.

5. Използваната литература отразява в голямата си част повечето основни разработки в областта, но може да се разгледат и допълнителни литературни източници, засягащи проблемите споменати в предходната точка. Два от литературните източници ([21] и [22]) в библиографията не са цитирани в текста на дисертацията. Описанието на библиографията трябва да бъде унифицирано като стил. На някои цитирани източници има непълно библиографско описание, например [39]
6. Бих искала да препоръчам на докторанта да публикува самостоятелни публикации в научно направление 4.6 Информатика и компютърни науки, с цел популяризиране на резултатите и получаване на обективна оценка на постиженията му в тази област от международни рецензенти.

7. Заключение

След като се запознах с представените в процедурата дисертационен труд и придружаващите го научни трудове и въз основа на направения анализ на тяхната значимост и съдържащи се в тях научни и научно-приложни приноси, **потвърждавам**, че представеният дисертационен труд и научните публикации към него, както и качеството и оригиналността на представените в тях резултати и постижения, отговарят на изискванията на ЗРАСРБ, Правилника за приложението му и съответния Правилник на СУ „Св. Климент Охридски“ за придобиване от кандидата на образователната и научна степен „доктор“ в научната област 4. Природни науки, математика и информатика и професионално направление 4.6. Информатика и компютърни науки. В частност кандидатът удовлетворява минималните национални изисквания в професионалното направление и не е установено плагиатство в представените по конкурса научни трудове.

Въз основа на гореизложеното, **препоръчвам** на научното жури да присъди на Илиян Недков Михайлов образователна и научна степен „доктор“ в научна област 4. Природни науки, математика и информатика, професионално направление 4.6. Информатика и компютърни науки.

23.08 2021 г.

Изготвил рецензията:

(доц. д-р Светла Бойчева)