Рецензия

по процедура за защита на дисертационен труд на тема:

„Интелигентни информационни системи в биоинформатиката:

семантично интегриране, анализ и класификация на био-медицински 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)

Рецензията е изготвена от: проф. д-р Евгений Христов Кръстев от СУ, ФМИ, катедра „Мехатроника, роботика и механика“ в качеството ми на член на Научното жури, съгласно Заповед № РД-38-255/02.06.2021 г. на Ректора на СУ. Заключенията в рецензията отчитат изискванията на следните нормативни документи:

1. Закон за развитието на академичния състав в Република България (ЗРАСРБ).
2. Правилник за прилагане на Закона за развитието на академичния състав в Република България (ППЗРАСРБ).
3. Правилник за устройството и дейността на СУ.
4. Правилник за условията и реда за придобиване на научни степени и заемане на академични длъжности в СУ (ПУРПНСЗАД).
5. Правилник за условията и реда за придобиване на научни степени и заемане на академични длъжности във ФМИ на СУ.

1. Обща характеристика на дисертационния труд и представените материали

Дисертационният труд на Илиян Недков Михайлов съдържа186 страници и се състои от 5 глави и Заключение, както и списък с използвана литература от 151 заглавия, списък на 50 фигури, списък на 16 таблици, речник на термините и списък на съкращенията. Докторантът е представил също така Автореферат от 37 страници, представящи обобщено съдържанието и характеристика резултатите от извършената научно- изследователска работа, както и списък с негови 9 публикации и един патент по темата на дисертацията.

Всички от останалите документи за провеждане на защитата на дисертационния труд, изисквани от ЗРАСРБ, Правилника за прилагане на ЗРАСРБ, ПУРПНСЗАД на СУ и ПУРПНСЗАД на ФМИ на СУ, са представени и оформени коректно от докторанта.

2. Данни и лични впечатления за кандидата

Докторантът завършва висше образование в ТУ Габрово през 2015 г. Придобива образователно-квалификационна степен "инженер-магистър" ТУ Габрово през 2018 г. в областта на компютърните системи и технологии. Илиян Недков Михайлов е зачислен като редовен докторант към катедра „Информационни технологии" през февруари 2018 г. и отчислен с право на защита със заповед РД 20-376/12.02.2021 на Ректора на СУ. От месец април 2017 г. до днес е Старши софтуерен разработчик в SAP Labs България, където специализира в разработването на софтуерни архитектури и разпределени системи в областта на облачните технологии. Според мен, докторантът прилага умело придобитите на работното си място знания и умения при изпълнение на научно-изследователските задачи в дисертационния труд. Организирал е семинар по приложения на бази от данни в биоинформатиката през 2020 г., участвал е с доклади в международни конференции по изчислителна биология, изкуствен интелект и компютърни науки.

Познавам Илиян Михайлов повече от 4 години във връзка участието му като хоноруван преподавател в курсове в магистърска програма Биомедицинска информатика , както и от съвместната ни работа по Национална Научна Програма Електронно здравеопазване в България. Това ми позволява да имам непосредствени впечатления от личните му качества и работата му по дисертационния труд. Винаги е бил сериозен и крайно отговорен при изпълнение на задачите си. Студентите във водените от него курсове имат положително мнение за работата му. Има мислене на изследовател, генерира и реализира нови идеи, има и способност да работи добре в екип. Постигнатите от него резултати ме убеждават във високата му професионална подготовка и умения в областта на информационните технологии и тяхното приложение в биоинформатиката. Напълно удовлетворява личните качества, необходими за придобиване на научно-образователната степен „доктор" в областта на професионалните си интереси.

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

В Глава 1“*Увод*“ е обоснована актуалността и значимостта на разглежданата проблематика. Формулирана е основната цел на дисертационния труд, разработването на автоматизирани и ефективни начини за интегриране на големи, хетерогенни набори от био- и медицински данни от множество източници. Като основно предизвикателство е посочена хетерогенността на източниците на данни, които са географски разпределени и разнородни по отношение на техните функции, структури, методи за достъп до данни и формати за разпространение. В тази глава са формулирани шест основни групи от проблеми във връзка със семантично интегриране, анализ и класификация на био-медицински данни. Такива проблеми възникват при съвместно използване на данни с клиничен, лабораторен и молекулярен профил, създаването на система за вземане на решения при извънклинична терапия на заболяване от диабет, класификация на метагеномни данни по микробиомна резистентност, приложение на методи за машинно самообучение за прогнозиране на структурата на протеини, компресиране ва данни от паралелно секвениране. Считам за много съществен елемент включването на задачата за разработване на платформа за предоставяне на софтуерните решения на тези проблеми под формата на услуги. Добро впечатление ми направи визуализацията на структурата на дисертационния труд на Фиг. 1.2, където е направена връзка между отделните глави и изпълнението на задачите, формулирани в първа глава.

Във втората глава (“*Теоретични основи и анализ на състоянието по проблемите за интегриране, анализи класификация на био-медицински данни*”) е направен литературен обзор на публикациите, свързани с работата посветени на модели за семантично интегриране на био-медицински данни на примера на данни от ракови заболявания: пациентски, клинични, -омикс данни. Това е една изключително интердисциплинарна област на научни изследвания, която изисква задълбочени познания в няколко различни научни направления. В литературния обзор е акцентирано на публикации върху съхранението и трансфера на данни, тяхното управление с приложение на нерелационни бази от данни, семантично интегриране на биомедицински данни и техния анализ. Семантичното интегриране на биомедицински данни е представено от модели и информационни технологии, ориентирани към интеграция на услуги. Не откривам ясно изразено разграничение между „*интегриране*“ и „*семантично интегриране*“(стр. 9) на данни. Считам, че за пълнота в този обзор щеше да е добре да се разгледат стандарти като например, *Observational*

Medical Outcomes Partnership (OMOP) Common Data Model (CDM), който дефинира широко разпространен и общоприет модел и процес за извличане, представяне и семантично интегриране на медицински данни, генерирани от хетерогенни бази от данни, посредством стандартни терминологии, речници и схеми за кодиране.

Третата глава „*Формализация и методи за интелигентно интегриране, анализ и класификация на биомедицински данни*“ разглежда създадените от докторанта модели и методи за интелигентно интегриране, анализи класификация на био-медицински данни. Представени са модели за интегриране на данни от множество различни източници, методология за семантично интегриране, използване на машинно обучение за оценка на базата на знания, извличане на знания от семантично интегриране и др. Практически интерес представляват резултатите, постигнати при интегриране на данни за микробиомна резистентност и последващия им анализ и класификация на разнообразие и разпространение. При изграждането на моделите са използвани средства на приложната математика и класическите математически дисциплини като изкуствен интелект и машинно обучение, теория на графи и дискретна математика. Тук се открояват оригинални методологични идеи на докторанта Илиян Михайлов, чиито реализации имат съществени теоретични и приложни приноси (Глава 5.1.1- 5.1.2). Разработени са модел за интегриране на хетерогенни биомедицински данни, модел ва съветваща система в полза на болните от диабет, модел за класификация и анализ на антимикробна резистентност и др.. Удачно за целите на изследването е избран нерелационен модел за обединяване на медицински данни. Предложена е клинична характеристика, обединяваща туморния стадий, размера на тумора и възрастта при поставяне на диагнозата, за оценка на преживяемост от раково заболяване със средствата на машинно самообучение. Същевременно, на докторанта са били нужни повече усилия, за да се изясни проблема за семантично интегриране на данни въз основа на стандартни терминологии, речници и схеми за кодиране.

В Глава 4 “*Софтуерна реализация на интелигентни системи за интегриране, анализ и класификация на био-медицински данни*” са представени резултатите от софтуерната реализация на разработените модели и използвани подходи в дисертацията, представени в предходната глава. Прави добро впечатление, че тези резултати демонстрират умения за работа с големи масиви данни, съхранявани, управлявани и интегрирани в нетрадиционни системи бази данни като например, NoSQL бази от данни. Създадена е архитектура на модерна хардуерна платформа за сложни разпределени изчисления, предназначена да изпълнява софтуерни решения като тези описани в дисертационния труд

под формата на услуги. Съществен елемент в реализацията на тази платформа и приложението на патент [C1], разработен в съавторство с докторанта. Там подчертано личат несъмнените знания в областта на информатиката, езици и технологии за програмиране на докторанта. Повечето от резултатите са получени със средствата на машинно самообучение.

Глава 5 „*Приноси и перспективи*“ изброява приносите на докторанта в дисертационния труд, перспективите за бъдещо развитие и документални справки в тази връзка. Основните приноси на докторанта са в областта на информационните технологии и компютърните науки във връзка с приложенията им в биоинформатиката (5.1.1- 5.1.2). Важно достойнство на получените резултати е, че те могат да намерят приложение с друг тип данни извън биомедицинската информатика. Това важи особено за разработената единна платформа за предоставяне на услуги.

В специално обособено *Заключение* се обобщават резултатите от изпълнението на задачите, формулирани в първата глава. Едновременно с това се дискутират и потенциални ограничения на предложените модели и тяхната реализация.

4. Апробация на резултатите

Докторантът е представил 9 публикации и 1 патент, регистриран(в съавторство) в САЩ, във връзка с темата на дисертацията, което значително надвишава необходимия минимум. Всички публикации са в съавторство с трима и повече автори. Поради липса на други данни приемам, че всички съавтори имат еднакъв принос. В четири от публикациите, докторантът е водещ автор ([C2, C3, C4, C7]).

Приносите в дисертационния труд произтичат от научните публикации на докторанта. Тези публикации са рецензирани и докладвани на 11 специализирани международни и национални конференции.

Всички публикации с изключение на [C9] и патента [C1] са в издания, реферирани от Scopus, Web of Science и IEEE Xplore, 7 от тях са с импакт фактор и/или SJR, към днешна дата тези публикации имат впечатляващ брой от **47 цитирания** ([C2]- 26, [C3]- 3 , [C4]- 14, [C6]- 1, [C7]- 3) и h-индекс = 3.

Научните трудове отговарят на минималните национални изисквания и значително надхвърлят тези 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. Не намирам данни за доказано по законоустановения ред плагиатство в представения дисертационен труд и научни трудове по тази процедура.

5. Качества на автореферата

Авторефератът съдържа 37 страници, отговаря на всички изисквания за изготвянето му и отразява коректно съдържанието на дисертационния труд.

6. Критични бележки и препоръки

Считам, че за по- доброто представяне на научно- изследователските резултати щеше да допринесе въвеждането на система от дефиниции на основни понятия (например, „*семантично интегриране на данни*“) и техни характеристики („*семантично сходство*“, „*семантично разстояние*“), около които да се изгради логическата структура на дисертационния труд.

При внимателен прочит на дисертацията могат да се открият редица неточности или непълноти при описанието на моделите и тяхната практическа реализация. Никъде в дисертацията не открих систематично описание на модела или схема на данните, които подлежат на семантично интегриране или просто, интегриране. Не са приведени дори примери на инстанции на такива данни. Същото се отнася и до данните, получени след завършване на процеса на интегриране. По тази причина е трудно да се формулират ограниченията в приложимостта и областите на валидност на резултатите.

В Глава 3.4 се предсказва „*точността на нагъване на протеинови структури*“ без да е посочено как се измерва (изчислява) тази точност. Същото се отнася и до точността на предвиждане на преживяемост при раково заболяване. В повечето случаи липсват или не са добре обосновани оценки на сходство от сравнителен анализ на получените числови резултатите по отношение на данни от реални наблюдения или резултати, налични в съществуващи публикации („*TICF показа по-добри резултати от NPI*.“, стр. 73). Формулите за статистическия модел на стр. 90-91 са изписани твърде небрежно. От въведените означения не се може да се разграничи кой параметър е индекс или вектор, какво е T, каква е интерпретацията на тези параметри и пр. Аналогични са проблемите на стр. 93 при описанието на „*оптимизиран алгоритъм*“. Използват се съкращения (ANN, RBF, стр. 74, LBFGS, стр 75) преди да се дефинират или еднакви съкращения (SVM, стр. 74), които имат различна интерпретация в дисертацията (SVM, стр. 107). Методологията по използване на моделите на машинно самообучение на стр. 74-75 има нужда от допълнително редактиране на текста („*обученият модел не е оскъден и по този начин е по-бавен от SVR, който научава оскъден модел за епсилон ??> 0, по време на прогнозиране*.“- стр. 75).

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, съдържащите се в тях научни и научно-приложни приноси, докладвани в **голям брой публикации и цитирания на тези публикации**, 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, напълно убедено **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** (Информационни технологии-Био и медицинска информатика).

20 август 2021 г. Изготвил рецензията: .....

проф. д-р Евгений Христов Кръстев