Review

of dissertation work:

„Intelligent information systems in bioinformatics: semantic integration, analysis and classification of biomedical data“

for covering of

the educational and scientific degree PhD

PhD candidate: **Ilian Nedkov Mihailov,**
Field of higher education: **4. Natural Sciences, Mathematics and Informatics,**
Professional Field**: 4.6 Informatics and Computer Science,**
PhD program: **„Information Technologies - Bio and Medical Informatics “,**
Department: **„Information Technologies“, Faculty of Mathematics and Informatics (FMI), Sofia University „St. Kliment Ohridski “ (SU).**
Supervisor: **Assoc. prof. Dimitar Ivanov Vassilev, PhD.**

Reviewer: **Assoc. Prof. Dr. Svetla Boytcheva, Institute of Information and Communication Technologies, Bulgarian Academy of Sciences,**
a member of the scientific jury for the competition according to Order No РД 38-255 /02.06.2021 of the Rector of SU.

1. **General description of the dissertation and the submitted documents**

   The dissertation is written in Bulgarian and has a volume of 186 pages: 165 pages main body and 21 pages including table of content, list of figures and tables, glossary of terms, references, and title pages. The text is organized in an introduction (Chapter 1), 4 chapters and a conclusion, and it should be noted that Chapter 5 contains the author's reference. The references list consists of 151 publications in English. Most of the cited publications have been published over the last 10 years. The references are appropriately cited in the dissertation text, with some minor exceptions. For example, for sources [21] and [22] there is no reference in the text of the dissertation.

2. Data and personal impressions of the PhD candidate

The PhD candidate Ilian Mihailov graduated in 2015 Bachelor's degree, specialty Computer Systems and Technologies, professional qualification "Computer Engineer" at the Technical University of Gabrovo. In 2017 he graduated with a Master's degree in Computer Systems and Technologies, professional qualification with a Master's degree in Computer Engineering at the Technical University of Gabrovo. From March 2014 to April 2017 he worked as a software developer at Bulpros. In the period 2018-2021 he was a doctoral student at the Faculty of Mathematics and Informatics of Sofia University "St. Kliment Ohridski", doctoral program: "Information Technologies - Bio and Medical Informatics", at the Department of Information Technologies. At the same time since April 2017, he has been working as a senior software developer in the software company SAP Labs Bulgaria.

I know the PhD candidate personally from his participation in scientific forums and I have excellent impressions of him as a responsible person and a well-established specialist in the field of computer science, who can do scientific research and present high-quality scientific results.

3. Content analysis of the scientific and scientific-applied achievements of the candidate, contained in the presented dissertation and the publications to it, included in the procedure

The dissertation aims to create a methodology and its practical implementation for intelligent integration of biomedical data and their analysis, using tools of informatics, information technology, bioinformatics, and artificial intelligence.

The **Introduction** (**Chapter 1**) section presents the motivation and discusses the importance of the research problem. The main goal of the dissertation and the tasks arising from it are defined, as well as the structure of the overall presentation is sketched.

Chapter 2 presents an overview, the theoretical basis and statement of the tasks are presented.

Chapter 3 describes data storage and integration models. The methods for horizontal and vertical integration of heterogeneous data are presented. A methodology for semantic integration of biomedical data from different diseases is presented. The application of linked data based on ontologies in the design of decision support in the healthcare domain is described. A new model for survival predicting cancer patients is presented. The application of machine learning methods in the evaluation of the accuracy of protein structures prediction is presented. A model for compressing omics data is presented. The application

of Gradient Boosting Machine, Random Forest, and Neural Networks in solving the problem of predicting antimicrobial resistance in metagenomic data is presented.

Chapter 4 provides a description of the software module for integrating heterogeneous data. The developed models for prediction, compression, and classification of different types of biomedical data are also described.

The author's reference is presented in **Chapter 5**.

The **Conclusion** section summarizes the presented scientific results and sets guidelines for future work.

The author's reference describes 5 scientific and 6 applied contributions. It would be good to reformulate, refine and fine-tune some of the scientific contributions' descriptions because in principle they sound very general and could lead to the misconception that solutions to the research problems are proposed in the general case and not in a specific too particular case, which is described in the dissertation.

4. Approbation of the results

From a formal point of view, in accordance with the Regulations for the implementation of the Act of the Development of the Academic Personnel in the Republic of Bulgaria (ADAPRB), Decree No 26 of 13.02.2019 for the amendment and supplement of Rules of implementation of ADAPRB (RIADAPRB), Field of higher education 4. Natural sciences, mathematics and informatics, Professional fields 4.1., 4.2., 4.3. , 4.4., 4.5., 4.6., and the Regulations for the Special Conditions for Acquisition of Academic Degrees and Academic Position in FMI-SU, the PhD candidate is required to have at least 30 points from criteria in Group G (indicators from 5 to 10).

The PhD candidate Ilian Mihailov has presented a total of 1 patent (C1) and 9 publications (C2-C10), as all publications are in English in peer-reviewed editions:

- Publication C9 is in a collection of abstracts and is excluded from consideration.
- Journal articles are as follows:
  - C2 (Biology Direct, Scopus-Q1, SJR 1.51; Web of Science-Q1, IF 4.54)
  - C4 (Information, MDPI, Scopus-Q3, SJR 0.353; Web of Science without quartile and without IF)
  - C8 (Frontiers in Genetics, Scopus-Q2, SJR 1.413; Web of Science-Q2, IF 4.599)
- Publications in conference proceedings are as follows:
  - C3 (Lecture Notes in Computer Science – Springer, Scopus-Q2, SJR 0.283; Web of Science-Q4, without IF)

- C5 (Lecture Notes in Computer Science – Springer, Scopus- Q3, SJR 0.249; Web of Science-Q4, without IF)
- C7 (Lecture Notes in Computer Science – Springer, Scopus-Q2, SJR 0.427; Web of Science-Q4, without IF)
- C6 (IEEE Explore, Scopus - without SJR)
- C10 (CEUR-WS, Scopus- without quartile, SJR 0.177)

All presented publications are co-authored with the research supervisor of the PhD candidate, as well as with other Bulgarian and foreign scientists. Ilian Mihailov is the first author of four of the publications (C2, C3, C4, C7). The PhD candidate has not submitted any individual publication. There are made in addition 11 presentations of the research results at scientific forums.

According to Decree № 26 of 13.02.2019 for the amendment and supplement of RIADAPRB are considered only scientific publications that are referenced and indexed in world-leading databases with scientific information (Web of Science and Scopus, Zentralblatt, MathSciNet, ACM Digital Library, IEEE Xplore and AIS eLibrary)

Thus, for point D (indicators 5 to 10) only 8 of the publications (C2-C8, C10) of Ilian Mihailov are considered. Then the total number of points on this indicator covers and many times exceeds the minimum requirement of 30 points.

There are 33 citations in Scopus of the PhD candidate's publications presented in the dissertation, excluding self-citations. The publications of the PhD candidate in reputable international scientific journals, as well as their numerous citations by world scientists, show the importance of the results obtained from scientific research in the field.

The presented scientific papers under this procedure meet the minimum national requirements (under Art. 2b, para. 2 and 3 of the Law on the Protection of Human Rights and Freedoms) and respectively the additional requirements of Sofia University "St. Kliment Ohridski" for acquiring the educational and scientific degree "Doctor" in the scientific field and professional direction of the procedure.

The presented results present original scientific work and do not repeat existing research presented in other procedures for obtaining scientific degrees and holding academic positions.

Data for plagiarism check have been provided and there is no legally proven plagiarism in the submitted dissertation and scientific papers under this procedure.

In my opinion, the PhD candidate's contribution to collective publications is clear and significant. It is evident from the subject matter. The doctoral student is the first author of four of the publications.

5. PhD Thesis summary

The presented PhD Thesis summary both in Bulgarian and English contains 37pages and reflects the main chapters and results of the dissertation. It is worth mentioning the difference in spelling the PhD candidate name in the PhD Thesis summary in English and the publications. This could lead to some confusion, and some unification will be helpful for further references.

6. Critical notes

I would like to make some technical remarks that do not reduce the value of the presented dissertation in terms of the results achieved by the PhD candidate, but would help to improve the presentation:

1. In my opinion, the main goal of the dissertation is not very well defined - it needs some refinement in the direction of emphasis on the research problem to be solved, as well as defining the subject and object of research, defining a working hypothesis from which to derive tasks of the dissertation. All defined tasks are too ambitious and cover many unresolved issues in the field of biomedical informatics, but there is no relationship between them with such a generally defined goal. The only link between the various subtasks is the field of application - biomedical informatics. This is one of the reasons why the results obtained are not in-depth research on a problem but are scattered in a very wide range of topics.
2. The text of the dissertation contains too many syntactic and grammatical errors. The use of slang words and foreign words, typically used in information technologies, is abundant. Despite the addition of a dictionary of the terms used, unfortunately, some of the terms are left there without translation into Bulgarian. It is also noteworthy that for some of the terms an incorrect translation into Bulgarian was used, for example for the main term "graph" in some places the term "graphics" was used, which is quite difficult to read and recognize in a particular context. Clearing the text of errors and improving the style of presentation would help to increase the quality of the presented dissertation.
3. The exposition of the dissertation is a bit chaotic and deviates from the standard format - a state-of-the-art of methods in the domain, a description of the proposed approach,

and an illustration and validation of the proposed approach application. Mainly the presentation focuses on what could be done in principle to solve the problem, but without sufficient specificity on how exactly this is done in the proposed solution. In general, the emphasis is more on the software components and implementation used than on the description of the proposed methods and algorithmic solutions. There are also not enough specific examples to illustrate the proposed approaches. In machine self-learning methods, experimental results with benchmark datasets are usually presented in order to make an objective assessment of the improvement of the proposed approach. It is difficult to distinguish the overview description of the state-of-the-art methods and the personal contribution of the PhD candidate in the exposition of the dissertation.

4. Regarding the proposed approach for the integration of heterogeneous data, three main issues are not addressed: (1) application of different methods for entity matching in the integration of data from heterogeneous sources; (2) concepts normalization of data into standard classifications and ontologies; (3) methods for information extraction – name entity recognition and relation extraction in the processing of semi-structured data, which contain text fields. The first problem is in particular related to the second - how to match entities from two databases, which are normalized through different ontologies. The third problem involves the use of AI approaches to Natural Language Processing, such as Information Extraction and text-based classification methods. These research problems are not solved in the general case, but they should be discussed in the dissertation and should be provided a more objective assessment of the specificity and the innovativeness of the proposed solution in the dissertation. Refining the PhD thesis description with a clear explanation of the limitations of the proposed approach for the integration of heterogeneous data would be useful to identify potential applications.
5. The references list reflects the major developments in the field, but some additional literature sources addressing the issues mentioned in the previous point can be considered. Two of the literature sources ([21] and [22]) in the bibliography are not cited in the text of the dissertation. The description of the bibliography should be unified as a style. Some cited sources have an incomplete bibliographic description, for example [39].
6. I would like to recommend to the PhD candidate to publish individual scientific publications in the field 4.6 Informatics and Computer Science, in order to promote the

results and to obtain an objective evaluation of his achievements in this field by international reviewers

7. Conclusion

After getting acquainted with the dissertation presented in the procedure and the accompanying scientific papers and based on the analysis of their significance and the scientific and scientific-applied contributions contained in them, I would like to **confirm** that the presented dissertation and scientific publications related to it, as well as the quality and originality of the results and achievements presented in them, meet the requirements of ADAPRB, the RIADAPRB and the respective Regulations of Sofia University "St. Kliment Ohridski" for acquisition by the PhD candidate of the educational and scientific degree "Doctor" in the field of higher education 4. Natural Sciences, Mathematics and Informatics, and professional field 4.6. Informatics and Computer Science. In particular, the candidate satisfies the minimum national requirements in the professional field for awarding the PhD and no plagiarism has been established in the scientific papers submitted at the competition.

Based on the above, I would like to **recommend** the scientific jury to award Ilian Nedkov Mihailov educational and scientific degree "Doctor" in the field of higher education 4. Natural Sciences, Mathematics and Informatics, Professional field 4.6. Informatics and Computer Science.

23.08 2021 г. Reviewer:

(Assoc. Prof. Svetla Boytcheva, PhD)