Review

on the procedure for defense of a dissertation on the topic:

"Intelligent information systems in bioinformatics:

semantic integration, analysis and classification of biomedical data "

to acquire educational and scientific degree "doctor"

Candidate**: Iliyan Nedkov Mihaylov**

Field of higher education: **4. Natural Sciences, Mathematics and Informatics**

Professional area: **4.6. Informatics and Computer Science**

Doctoral program: **"Information Technology-Bio and Medical Informatics",**

Department**: Information Technology**

Faculty of Mathematics and Informatics (FMI),

Sofia University "St. Kliment Ohridski "(Sofia University)

The review is prepared by: Professor Dr. Evgeniy Hristov Krastev from Sofia University, FMI, Department of Mechatronics, Robotics and Mechanics in my capacity as a member of the Scientific Jury, according to Order № RD-38-255 / 02.06.2021 of the Rector of Sofia University. The conclusions in the review take into account the requirements of the following normative documents:

1. Academic Staff Development Act in Republic of Bulgaria (ZRASRB),
2. Regulations Act about the structure and activity of Sofia University
3. Regulations Act about the Implementation of ZRASRB (PPZRASRB)
4. Regulations Act about the Terms and Conditions for Acquisition of Academic Degrees and Occupation of Academic Positions at SU (PURPNSZADSU).
5. Regulations Act about the terms and conditions for acquiring scientific degrees and holding academic positions in the FMI of Sofia University

1. General characteristics of the dissertation and the presented materials

The dissertation of Iliyan Nedkov Mihaylov contains 186 pages and consists of 5 chapters and a Conclusion, as well as a list of references comprising 151 titles, a list of 50 figures, a list of 16 tables, a glossary of terms and a list of abbreviations. The doctoral student has also presented

an Abstract of 37 pages, summarizing the content and characteristics of the results of the research work, as well as a list of his 9 publications and a Patent on the topic of the dissertation.

All of the other documents for conducting the defense of the dissertation, required by ZRASRB, the Regulations for application of ZRASRB, PURPNSZAD of SU and PURPNSZAD of FMI of SU, are presented and prepared correctly by the doctoral student.

2. Biographical data and personal impressions of the candidate

The doctoral student graduated from the Technical University of Gabrovo in 2015. He earned the educational qualification degree "Engineer-Master of Science" from the Technical University of Gabrovo in 2018 in the field of computer systems and technologies. Iliyan Nedkov Mihaylov was enrolled as a full-time doctoral student at the Department of Information Technology in February 2018 and completed his studies with the right to defend a dissertation by order RD 20-376 / 12.02.2021 of the Rector of SU.

From April 2017 until today he is a Senior Software Developer at SAP Labs Bulgaria, where he specializes in the development of software architectures and distributed systems in the field of cloud technologies. In my opinion, the doctoral student skillfully applies the knowledge and skills acquired at work in performing research tasks in the dissertation. He has organized a seminar on database applications in bioinformatics in 2020, participated with reports in international conferences on computational biology, artificial intelligence and computer science.

I have known Iliyan Mihaylov for more than 4 years in connection with his participation as a part-time lecturer in courses in the master's program Biomedical Informatics, as well as from our joint work on the National Scientific Program eHealth in Bulgaria. This allows me to have immediate impressions of his personal qualities and his work on the dissertation. He has always been serious and extremely responsible in carrying out his tasks. The students in the courses he teaches express positive opinion about his work. The doctoral student has the thinking of a researcher, generates and implements new ideas, has the ability to work well in a team. The results achieved by him convince me of his high professional experience and skills in the field of information technology and their application in bioinformatics. Fully satisfies the personal qualities necessary for obtaining the scientific-educational degree "Doctor" in the field of his professional interests.

3. **Detailed 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**

Chapter 1 "*Introduction*" substantiates the relevance and importance of the issues under consideration. The main goal of the dissertation is formulated, the development of automated and effective ways to integrate large, heterogeneous sets of bio- and medical data from multiple sources. The main challenge is the heterogeneity of data sources, which are geographically distributed and heterogeneous in terms of their functions, structures, data access methods and distribution formats. This chapter sets out six main sets of issues related to the semantic integration, analysis, and classification of biomedical data. Such problems arise from the joint use of data with clinical, laboratory and molecular profile, the creation of a decision-making system for outpatient therapy of diabetes, classification of metagenomic data by microbiome resistance, application of machine self-learning methods to predict the structure of proteins, compressing data from parallel sequencing. I consider the inclusion of the task of developing a platform for providing software solutions to these problems in the form of services to be a very important element. A good impression makes the visualization of the structure of the dissertation in Fig. 1.2, where the individual chapters are related to the proposed solutions of the tasks assigned in Chapter 1.

In the second chapter ("*Theoretical foundations and analysis of the state of integration problems, analysis classification of bio-medical data*") is a literature review of publications related to the work on models for semantic integration of bio-medical data on the example of data from cancers: patient, clinical, -*omix* data. This is an extremely inter-disciplinary field of research that requires in-depth knowledge in several different scientific fields. The literature review focuses on publications about data storage and transfer, their management using non-relational databases, semantic integration of biomedical data and their analysis. The semantic integration of biomedical data is represented by models and information technologies oriented towards the integration of services. I do not find a clear distinction between "*integration*" and "*semantic integration*" (p. 9) of data. I believe that for completeness in this dissertation it would be good to consider standards such as the Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM), which defines a widely used and generally accepted model and process for extracting, presenting and semantically integrating medical data. generated by heterogeneous databases, using standard terminology, dictionaries and coding schemes.

The third chapter "*Formalization and methods for intelligent integration, analysis and classification of biomedical data*" discusses the models and methods created by the doctoral student for intelligent integration, analysis and classification of biomedical data. Models for data integration from many different sources, methodology for semantic integration, use of machine learning for knowledge-based assessment, knowledge extraction from semantic integration, etc. are presented. Of practical interest are the results achieved in the integration of microbiome resistance data and their subsequent analysis and classification of diversity and distribution. In the construction of the models, means of applied mathematics and classical mathematical disciplines such as artificial intelligence and machine learning, graph theory and discrete mathematics were used. Here stand out original methodological ideas of the doctoral student, whose realizations have significant theoretical and applied contributions (Chapter 5.1.1-5.1.2). A model for integration of heterogeneous biomedical data, a model of an advisory system in favor of patients with diabetes, a model for classification and analysis of antimicrobial resistance, etc. have been developed. A clinical property combining tumor stage, tumor size, and age at diagnosis has been proposed to assess cancer survival in terms of machine learning methods. At the same time, more efforts are needed to clarify the problem of semantic integration of data based on standard terminology, dictionaries and coding schemes.

Chapter 4 "*Software implementation of intelligent systems for integration, analysis and classification of bio-medical data*" presents the results of the software implementation of the developed models and approaches used in the dissertation, presented in the previous chapter. It is noteworthy that these results demonstrate skills for working with large arrays of data stored, managed and integrated into non-traditional database systems such as NoSQL databases. An architecture of a modern hardware platform for complex distributed computing has been created, designed to implement software solutions such as those described in the dissertation in the form of services. An essential element in the implementation of this platform and the application of a patent [C1], developed in co-authorship with the doctoral student. There, the doctoral student demonstrates undoubted knowledge in the field of informatics, languages and technologies for programming. Most of the results were obtained by means of machine learning tools.

Chapter 5 "*Contributions and Perspectives*" describes the doctoral student's contributions in the dissertation, potential for future development and documentary references in this regard. The main contributions of the doctoral student are in the field of information technology and

computer science in connection with their applications in bioinformatics (5.1.1-5.1.2). An important advantage of the obtained results is that their application can be extended with other types of data outside of biomedical informatics. This is especially true for the developed unified platform for providing services.

A separate section *Conclusion* of the dissertation summarizes the results of the implementation of the tasks formulated in the first chapter. At the same time, potential limitations of the proposed models and their implementation are outlined.

4. Approbation of the results

The doctoral student has presented 9 publications and 1 patent, registered (in co-authorship) in the USA, in connection with the topic of the dissertation, which significantly exceeds the required minimum. All publications are co-authored with three or more authors. Due to the lack of other data, I accept that all co-authors have the same contribution. In four of the publications, the doctoral student is a lead author ([C2, C3, C4, C7]).

The contributions to the dissertation work derive from the doctoral student's scientific publications. These publications have been reviewed and reported at 11 specialized international and national conferences.

All the publications except [C9] and the patent [C1] are in publications referenced by Scopus, Web of Science and IEEE Xplore, 7 of them are impact factor and / or SJR, to date these publications have an impressive number of 47 citations ([C2] - 26, [C3] - 3, [C4] - 14, [C6] - 1, [C7] - 3) and h-index = 3.

The scientific works meet the minimum national requirements and significantly exceed these requirements (under Art. 2b, para. 2 and 3 of ZRASRB) 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. I do not find data for legally proven plagiarism in the submitted dissertation and scientific papers on this procedure

5. Evaluation of the abstract

The abstract contains 37 pages, meets all the requirements for its preparation and correctly reflects the content of the dissertation.

6. Critical remarks and recommendations

I believe that the introduction of a system of definitions of basic concepts (eg "*semantic integration of data*") and their characteristics ("*semantic similarity*", *"semantic distance*") would have contributed to the better presentation of research results. to build the logical structure of the dissertation.

Upon careful reading of the dissertation, a number of inaccuracies or incompleteness can be found in the description of the models and their practical implementation. Nowhere in the dissertation did I find a systematic description of the model or scheme of data that is subject to semantic integration or simply, integration. Not even examples of instances of such data are given. The same applies to the data obtained after the completion of the integration process. For this reason, it is difficult to formulate limitations in the applicability and areas of validity of the results.

Chapter 3.4 predicts the "*folding accuracy of protein structures*" without specifying how this accuracy is measured (calculated). The same goes for the accuracy of predicting cancer survival. In most cases, there are no or not well-founded estimates of similarity from a comparative analysis of the numerical results obtained in terms of data from actual observations or results available in existing publications ("*TICF showed better results than NPI*.", page 73) . The formulas for the statistical model on pages 90-91 are written rather carelessly. From the entered notations it is not possible to distinguish which parameter is an subindex or vector, what is T, what is the interpretation of these parameters, etc. There are similar problems on page 93 in describing an "*optimized algorithm*". Abbreviations (ANN, RBF, page 74, LBFGS, page 75) are used before defining either the same abbreviations (SVM, page 74), which have different interpretations in the dissertation (SVM, page 107). The methodology for using machine self-learning models on pages 74-75 needs further text editing (“*the trained model is not scarce and thus slower than the SVR, which learns a scarce model for epsilon ??> 0, during forecasting*. ”, page 75).

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, the scientific and scientific-applied contributions contained in them, reported in **a large number of publications with numerous citations of these publications**, I confirm that the **presented dissertation and**

scientific publications to it, as well as the **quality and originality** of the results and achievements presented in them, **meet the requirements** of ZRASRB, the Regulations for its application and the respective Regulations of Sofia University "St. Kliment Ohridski "for acquisition by the candidate of the educational and scientific degree" Doctor "in the scientific field 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 and no plagiarism has been established in the scientific papers submitted at the competition.

Based on the above, **I strongly recommend** the Scientific Jury to award **Iliyan Nedkov Mihaylov** the educational and scientific degree "Doctor" in scientific field **4. Natural Sciences, Mathematics and Informatics**, professional field **4.6. Informatics and computer science** (Information technology-Bio and medical informatics).

August 20, 2021

Review prepared by:

Professor E. Krastev, PhD