REVIEW

by Prof. Galia Angelova,
Institute of Information and Communication Technologies (IICT),
Bulgarian Academy of Sciences (BAS)

for the thesis of Kristina Ivanova Dineva

"*Integration of Heterogeneous Data from Distributed IoT Devices*"
presented for awarding the educational and scientific degree "PhD"
in professional area 4.6 "Informatics and Computer Sciences"

Pursuant to Order 211/2.11.2020 of the Director of IICT-BAS, I am nominated as a member of the Scientific Committee for the defense of the submitted thesis "Integration of Heterogeneous Data from Distributed IoT Devices" in professional area 4.6 "Informatics and Computer Sciences". The candidate Kristina Dineva has been a full-time doctoral student at IICT-BAS in 2017-2019. The dissertation is related to some of the present challenges in Information and Communication Technologies: collecting big data from different devices for the purpose of observing objects and processes; integrating large arrays of heterogeneous data and processing them (most often by machine learning) for the needs of practical applications; implementation of real-time systems using Internet of Things (IoT) paradigm. These topics are among the hottest trends in the emerging data industry and a subject of greatest interest from both researchers around the world and business. The application domain chosen for experiments in this dissertation is "intelligent agriculture" - one of the foci in a number of Bulgarian strategic documents for innovative development and implementation of digital technologies in the next decade. So the first positive impression about the thesis comes from the ambitious choice of a hot contemporary topic, which will develop further over a relatively long period in the future.

Formally, according to the Rules for the Implementation of the Law on the Development of the Academic Staff in the Republic of Bulgaria of July 6, 2018, a PhD candidate in 4.6 "Informatics and Computer Science" is expected to meet the minimum requirements of 50 points for indicator A (presentation of doctoral thesis) and to have minimally at least 30 points in the "Г"-group of indicators. Kristina Dineva presents a list of 10 scientific articles in which she is a co-author, and two articles in Bulgarian she is the only author. The articles were published in the period 2017-2019 in connection with the results of the dissertation, 8 of them are indexed by Scopus and 6 have an SJR-rank, with one paper in Scopus Q2 quartile. A list of 19 citations of the published papers is presented too, and 13 of them are in scientific articles by foreign authors outside the professional team where the candidate Ms. Dineva works. The points collected from the publishing activity (190) and the citations exceed the requirements about a minimal threshold for the indicators group "Г" in case of PhD dissertation defence. With this the formal conditions for satisfying the national criteria are fulfilled and it is possible to start reviewing and defending the presented dissertation.

The aim of the PhD thesis is "to offer a system and tools for integrating heterogeneous data from distributed IoT devices to allow for their processing, modeling and integration". The following tasks are set: to propose a methodology for processing, modeling and integration of heterogeneous data; to propose an architecture and method for communication in a modular IoT hardware system, as well as an architecture of a software platform and an approach for the organisation of services for intelligent processing of heterogeneous data from IoT systems; to create well-trained machine learning models for experimental validation and to show possible useful applications of the resulting system.

The dissertation contains 166 pages of text which is organised in an introduction, five chapters and a conclusion. Some 175 references are cited. The thesis is illustrated with 55 figures and contains 16 tables. A glossary of terms and abbreviations is included, which facilitates the understanding of abbreviations in English.

Content and contributions of the thesis

Chapter 1 presents related approaches and analyses the state of the art. Due to the wide range of tasks close to the topics under consideration, short overviews are given concerning Industry 4.0, IoT Internet of Things (including hardware, software, communications, platforms), IoT applications, IoT challenges, heterogeneous data and challenges for their collection and maintenance, solutions for integration of heterogeneous data. The conclusion is that it is necessary to create a unified and systematic approach for the integration and processing of heterogeneous data.

Chapter 2 proposes a methodology for processing, modeling and integration of heterogeneous data. A strategy for organising the workflow in four stages is shown (problem definition; data preparation - data collection, cleaning, transformations, data analysis; task modeling through machine learning - algorithm selection, configuration, training and validation; integration of the trained model in an application environment). At a first glance, for any scientist with many years of experience in artificial intelligence, the described steps for data preparation and processing might look common, but the presented conceptual scheme has certain merits. It is characterised by a clear distinction of the individual stages and by decomposing the task into smaller modules, for which solutions based on different technologies are offered. This chapter shows the careful preparation of the PhD candidate to identify a variety of approaches, algorithms, technologies and ready-made software tools which might be suitable for the implementation of the various stages of data processing and model creation. This chapter of the dissertation reflects the knowledge acquired by Kristina Dineva in the field of "data science" and the experience gained with different data processing environments.

Chapter 3 discusses the implemented system for data collection, processing and modeling. The system consists of remote IoT devices (installed in beehives) and a distributed cloud-based

software application that provides monitoring of data processing results through user interfaces. Data are collected in real time or at regular intervals. The communication between the hardware components is wireless, a router is located in the center of the system. The architecture of the software system is adaptive, currently based on eight types of microservices with the possibility of expansion. The user interface implemented through the Angular platform allows to users in the descriptive section to enter and update information about their apiaries and hives, and in the control section – to monitor in real-time the condition in the hives (for those users who have installed the IoT system in their apiaries). This chapter demonstrates the skills of Kristina Dineva as a professional computer scientist and expert in information and communication technologies, who is capable to develop hardware and software components, offers original IP-addressing scenarios and solutions related to energy efficiency of the system. The implemented system is an effective prototype, prepared for real application, and thanks to its modularity allows for the inclusion of additional devices and new functionalities.

Chapter 4 presents an experiment with a real practical application in order to validate the proposed methodology for processing and integration of heterogeneous data obtained from IoT devices. Data collected from observations of honey bees is used to predict two situations: "the bee family does not have sufficient conditions for survival" or "the bee family has sufficient conditions for survival". A binomial classifier has been trained over data indicating indoor and outdoor temperatures of beehives and atmospheric conditions, and comparative studies of the model's success have been performed with four classification algorithms of Microsoft Machine Learning Azure Studio cloud. Results for accuracy, precision, and sensitivity of the trained model were obtained, with a harmonic mean of 0.845. The relationships and correlations between the measured indicators/variables were studied by regression analysis. It has been shown that the most important factors with different coefficients of influence on the change in the amount of honey in a beehive are: temperature, atmospheric pressure, humidity, wind speed and time. Estimates are made of expected bee quantities within 10 days in order to facilitate the activity planning by beekeepers who have to visit their apiaries. The approach and results of this experiment are impressive with its completeness, depth and comprehensiveness of data analysis, as well as with its orientation to the needs of the users in a real application.

Chapter 5 presents the practical application of the implemented IoT system called SmartBeeHives. The system combines "all main components of ICT" - software, hardware, transactions, communications, data, Internet access and cloud computing. A complete service for the end user is provided – data collection, transmission, processing, analysis, modeling, data visualization and event prediction. The field of application is the so-called "precision beekeeping", where digitalisation enters mostly as a technology for remote monitoring and control. This chapter presents also a brief overview and comparative analysis of four systems for intelligent beekeeping, which are compared with the candidate's system SmartBeeHives. The advantages of SmartBeeHives are shown. They are largely due to the modularity and flexibility of the system, which can be built on different (already existing) devices.

Discussion

I accept the author's scientific and applied contributions as declared in the conclusion of the dissertation. However, for me the greatest achievement of this thesis is the ambitious approach of Kristina Dineva to start with basic theoretical considerations and in 3-4 years to reach a product prepared for marketing. The in-depth data analysis, based on sound statistical methods, is a guarantee for calculating the best possible prediction of the processes taking place in the hives, in the context of a relatively simple IoT system. In addition, it is possible to repeatedly train the SmartBeeHives's predictive core with new data collections and to update the system by improved functionality without interrupting the operation of the software environment. The conceptual simplicity of the hardware structure and its flexibility are an evidence for the author's expertise in the development of modern ICT-based applications.

The thesis text is concise and specific, appropriately organised with a clear division into chapters and topics, with a reader-friendly sequence of the presentation, sufficient level of detail and well-illustrated with figures and tables. The inclusion of a list of author's publications after each chapter is useful because it helps to understand where the results are presented. The Abstract correctly reflects the content of the dissertation.

I would like to address here the issue of terminology translation: for example, "машинно самообучение" is used in Bulgarian (this is the title of the corresponding article on "Machine Learning" in the Bulgarian Wikipedia), and the English term "confusion matrix" / "matrix of errors" is usually translated as "матрица на грешките". Another, more conceptual question is whether the numerical data considered in the experiment are "truly" heterogeneous, but I do not think it is the right place to discuss here the definitions of heterogeneity in the IoT field. Anyway, the methodology for data integration and processing proposed in chapter 2 of the thesis would look principally in a similar way for scenarios with much more complex varieties of data with different nature.

Conclusion

The Law on the Development of the Academic Staff in the Republic of Bulgaria requires any PhD thesis to contain scientific or applied results which represent an original contribution to the respective scientific discipline. The thesis should show that the candidate possesses deep theoretical knowledge in the corresponding scientific field and abilities for doing independent research. These requirements are fully met in the case of Kristina Dineva's dissertation and the scientific publications where the thesis results are presented. The relatively large number of citations is an evidence of the international interest and the merits of the thesis. The dissertation impresses with its ambition to produce a real practical application of high quality, which can be positioned at the world level and compared to the best systems in the field of precision beekeeping.

On these grounds, I will vote positively to award the degree and I propose with conviction to the honourable **Scientific Committee to award to Kristina Dineva the educational and scientific degree "doctor" (PhD) in computer science**.

30 November 2020

Member of the Scientific Committee

Prof. Dr.Sc. G