Opinion

on the dissertation thesis of Veselin Lazarov Shahpazov titled: „Research of models for forecasting capital markets with neural networks" in Doctoral Programme „Informatics", Professional Area 4.6. „Informatics and Computer Sciences"

by Assoc. Prof. Vassia Krassimirova Atanassova, PhD,
from Bioinformatics and Mathematical Modelling Department,
Institute of Biophysics and Biomedical Engineering, Bulgarian Academy of Sciences

By Order 262 from 31 October 2019 of the Director of IICT-BAS, I was appointed as a member of the Scientific Jury in the procedure for awarding the educational and scientific degree "Doctor of Philosophy" in the Professional Area 4.6. "Informatics and Computer Science", Doctoral Program 01.01.12 "Informatics" to Veselin Lazarov Shahpazov. On 27 November 2019 I received in electronic form the documents for the procedure, namely: a dissertation thesis, a synopsis in Bulgarian, a synopsis in English, a CV, a Reference for compliance with the minimal requirements of IICT-BAS and copies of 6 publications. In the short timeframe set at the first meeting of the Jury on 4 November 2019, which I was unable to attend due to conference participation, I was able to get acquainted with the dissertation thesis and the synopsis in Bulgarian, as well as with the publications of the student, but I had no opportunity to review the English version of the synopsis, hence I will not comment on the latter.

My impressions of the material presented herewith are mixed. On the one hand, part of the publications make quite a good impression, especially the self-written article "High Frequency Trading" in Bulgarian in the "Technosphere" popular science magazine from 2016, and the well-cited paper "Design and Application of Artificial Neural Networks for Predicting the Values of Indexes on the Bulgarian Stock Market" presented at SPS'2013. On the other hand, the dissertation thesis and the synopsis based on these publications leave the impression that they had been prepared in a hasty manner, without the necessary attention to detail. What further makes an impression is that the experimental work was based on data from the period 2010-2013.

With the exception of the recent report at a conference in Plovdiv, which is in press, all other publications on the dissertation thesis are from the period 2013-2016, and with the exception of the one in the "Technosphere" magazine, which is in Bulgarian, all others are communications at research forums and are written in co-authorship. A publication in a peer-reviewed scientific journal, with at least SJR rank, would have been an even stronger attestation, but the data provided in the Reference for compliance with the minimal requirements of IICT-BAS for Doctorate in Professional Area 4.6 "Informatics and Computer Sciences" regarding Veselin Shahpazov show that he meets the minimal requirements stipulated in the Regulations of IICT-BAS.

Considering the thesis and the synopsis, it would be difficult – and unnecessary – to list all the inaccuracies of terminological and factual nature, abundantly exhibited herein. I will mention only some of them:

- On page 8: wrong translations of *„genetic algorithms... ant colony optimization"*.
- On page 15: wrong translations of the term "time series".
- On page 26: „The Age of Camelot" has been defined by Eberhart and Dobbins as the time interval 1890 – 1969, rather than 1890 – 1926, as written in the thesis, which can be verified with the abstract available at https://ieeexplore.ieee.org/document/59207).
- On page 30, it is stated that *„Research in this field has grown exponentially over time"*, which, apart from being a frivolously picked idiom, does not correspond to the literary review on pages 81-82 and to the author's personal opinion stated on page 13 about the *„focus on the niche of unexplored and little-known markets like the Bulgarian"*, and *„little research in the area per the author's humble opinion"*.
- On pages 35, 61, 75, and 81, the term *"feedforward neural network"*, which does have an established translation in Bulgarian, is found in various extravagant variations.
- On page 38, formula (3) is wrong from mathematical point of view: initially f is binary, and later becomes a unary function. The way formula (6) is described on page 40 shows that λ is a function, not a function's parameter as it has been previously declared.
- On page 39, formulas (8) and (9) are inconsistent, as the parameter α is dropped out somewhere between the two of them.
- On page 43. the interval [–1; 1] is separated between the lines in a completely impermissible way.
- On page 48 (and elsewhere), the term "radial basis function network" is translated in a variety of wrong ways, and only at twice throughout the thesis has it been correctly guessed.
- Again on page 48, there is an unintelligible sentence *„Although the network ..."*.
- On page 52, it is stated that *„The Least Squares algorithm is proposed in 1960 by Widrow and Hoff, also known as the Widrow-Hoff rule [65]"*. This statement is incorrect: the Least Squares algorithm was proposed by Carl Friedrich Gauss in 1795. The concept proposed by Widrow and Hoff is the *Least Mean Squares*. While sounding similarly in both English and Bulgarian, the two concepts are charged with different mathematical meanings, and mixing them is impermissible.
- On page 61, it is written *„where Y is the output of the neural network"*, but nowhere in the concrete text there is a symbol *Y*.
- Many times within the text, e.g. on pages 72, 81, 85, and 86, one can come across the expressions *„stop at the digit 20"*, *„where the digit is 19"*, *„around the digit of 350"*, and *„For Bulgaria, the digit is around 4%."* Such examples of journalistic and pseudo-scientific vocabulary have no place in a dissertation thesis, least of all in the Scientific Area 4 „Mathematics and Informatics".

- On page 78: inaccurate Bulgarian translation of the term "naïve (Bayesian network/classifier) predictors".
- On pages 89-90, from formulas (29) and (30) it does not become clear how the first day (e.g. $t = 1$) is going to be computed.
- On pages 89, 90, and 94, the sub-indices in the text descriptions after the formulas have been inexplicably neglected.
- On page 94 and elsewhere, a number of different wrong translations of the term "multilayer perceptron" can be found.
- On page 107, a wrong variation of the established Bulgarian translation of the term "*curse of dimensionality*" is given.
- On page 111, a wrong translation of the term „(statistical) outliers" is given; two more precise Bulgarian versions have been proposed by the reviewer.
- Examples of meaningless fractions of sentences on pages 111, 112, and 114 have been listed by the reviewer.
- On pages 113, 115, and 117, an obvious machine translation of the term „k-means" can be seen.
- On page 118, the expression „network architectures" in the area of computer sciences is charged with a different meaning.

Apparently, some of the terminological inaccuracies have come as a result of using mainly English-language sources and adopting them in the dissertation thesis after (unedited) machine translation, despite the presence of a conceptual framework and sufficient resources on neural networks in Bulgarian, too. The probable explanation for some other mistakes is the applicant's insufficient fluency in the part of the problem area related to mathematical modelling with neural networks. I will not comment on the spelling and punctuation errors, although they make a bad impression not only given the level of the thesis presented – for the highest of the educational degrees – the educational and scientific degree of "Doctor of Philosophy", but also because such elementary mistakes often hinder the immediate comprehension of the text.

The structure of the thesis and the synopsis is another aspect which calls for serious comments and recommendations. In practice, the thesis features a literary review of capital markets models with artificial neural networks, which comprises about 80 of the approximately 130 pages of text. While I do find it informative and fascinating to read the brief presentation of the 53 research papers selected for analysis, I cannot fail to note that the review part is disproportionate to that of the original author's results, which are lost among the vast amount of ballast information of diverse nature. At the end of the literary analysis, item 2.7.2. on pages 80-82, both the text of the section and Figures 7–9 (which, in addition, repeat the information from the text in graphical form while not carrying additional semantics) should have cited the corresponding analysed 53 publications; nevertheless this weakness is offset by the generally informative Table 5.

In the course of the literary analysis, some of the conclusions of the authors of the analysed publications, which the PhD student has taken for granted, could and should have been critically evaluated in view of the time passed since these papers were

published, and revised accordingly. For example, I refer to statements as the one on page 74 regarding the cited dissertation thesis [49]. *"Authors' studies show that no such models are currently used in Bulgaria. One reason for this is the limited number of professionals in Artificial Intelligence and the lack of knowledge of the powerful technologies in this field."* This statement may have been valid as of 2008, but I certainly do not think it is valid in 2019, as the text suggests is the applicant's consideration.

Particularly chaotic is Chapter III, specifically Section 3.4, which lacks any logical link between the subsections whatsoever, and continuity of the presentation. In addition, in the introduction of the concept of expert system in item 3.4.3, apart from the errors in the translated terminology ("domain", "infinite loophole"), an example is given in Table 12, which is both unnecessary for the informed readers and uninformative for those who might find it necessary; it would make a lot more sense if, instead of low temperatures, slippery roads and red ears, an illustrative example with a capital market situation was considered.

In essence, it is difficult to formulate specific criticisms regarding the hybrid model proposed in this chapter and the results of the dissertation thesis. The author refers to the *"poorly researched Bulgarian market"* but declares to have taken into account (p. 98) *"the peculiarities of the Bulgarian stock market, and in particular its illiquidity, which allows for individual transactions, not executed on a purely market principle, to affect the overall index value and its change."*

Furthermore, as far as structuring is concerned, I believe that the introduction of Roman numerals for chapters and Arabic for – in fact – a single heading at first level within a chapter, besides being unnecessary, creates confusion for both the author and the reader. The numbering of formulas not referenced within the text is also unnecessary. Formulas that are not typed within the text but cropped as images from (unspecified) other sources can be seen at many instances, which, besides the student's ability to handle mathematical texts, raises the question whether the denotations used throughout the text have been validly and consistently introduced (and many a time the answer is – no).

In preparing the synopsis, the numbers of the formulas and figures in the dissertation thesis should be preserved to facilitate the traceability between the two documents. This would have prevented anomalies such as the appearance of "Figure X" on page 16 of the synopsis.

Many of the citations in the bibliography are incorrectly described.

In conclusion, I have noted all the aforementioned terminological inaccuracies for two reasons in particular. On the one hand, it is indisputable that the topic of the thesis is trendy, and it can be considered with high degree of confidence that the results obtained by the PhD student will be of interest to practitioners in the field of capital markets. On the other hand, according to the procedure of protection enshrined in the Law, the synopsis is given publicity and its content can in turn become a literary source for other doctoral students, scholars and practitioners in the field, who may use and cite it in the

future, and thus multiply in the literature the mistakes contained in this work. In this sense, while that may not be envisaged by procedure, I would sincerely recommend that the doctoral student revises at least the online version of his synopsis, taking into consideration as many of these critical remarks as possible.

Regarding the noticed citations of the publication *„Design and Application of Artificial Neural Networks for Predicting the Values of Indexes on the Bulgarian Stock Market"*, I will remark that I would not accept the citation

> Samit Bahnja, Abhishek Das, *Impact of Data Normalization on Deep Neural Network for Time Series Forecasting*, 2018.

which represents an unpublished preprint in arXiv, however, in my research I have discovered another citation of the same publication from 2014, available in SCOPUS:

> Shi, H., Liu, X. (2014). *Application on stock price prediction of Elman neural networks based on principal component analysis method*. 11th International Computer Conference on Wavelet Active Media Technology and Information Processing, ICCWAMTIP 2014, 7073438, pp. 411-414.

Given that there is no requirement for the PhD candidate to have any citations whatsoever, I highly commend the presence of these six citations of one of the papers, which the dissertation thesis has been based on.

Questions to the PhD student

- Explain the difference between the Least Squares method and the Least Mean Squares method?
- On page 82, it is written *„The challenge of making money from investments in financial instruments requires nevertheless advanced methods and correct selection of the various model components, such as the input data, the architecture of the neural network, the training methods and eventually the way of materializing the results produced."* While the reviewer principally agrees with this statement, the question to the PhD candidate is to specify the way he has selected the parameters of the neural networks' architectures (Table 8 on page 113 and Table 10 on page 118). How did he define the number of input elements, those in the hidden layer, and at the output of the three analysed types of ANNs? Of specific interest are the parameters of the general regression ANN.
- On page 94, it is written *„...in this case the online version of the back propagation algorithm has been used" and on page* 112 is written *„The software product used for their experimental implementation uses ..."* The PhD candidate is requested to specify what "online version" and which "software product" has been considered?
- On page 75 in the literary review, there is an analysis of publication [47] that is related to forecasting the prices of nine stock options and the Nifty 50 Index, before and after the 2016 Indian banknote demonetization. In the hypothetical case of launching the Euro in Bulgaria, would it be possible for the herewith proposed hybrid

model to forecast the two indices discussed in the thesis, and what would the results be?

Considering the fact that practically all the errors noted above can be possibly eliminated, i.e., these are not insurmountable weaknesses of this work; and considering the indisputable relevance of the topic of the dissertation thesis and the possibilities for the development of the thesis' problem area and the doctoral student himself, as well as considering the fact that the dissertation thesis meets the requirements of the Law and the applicable Regulations of IICT-BAS, I give a positive opinion and recommend to the honourable members of the Scientific Jury to award Veselin Shahpazov with the degree "Doctor in Philosophy" in Professional Area 4.6 "Informatics and Computer Science", Doctoral Program "Informatics".

3 December 2019

Author of the opinion:

Assoc. Prof. Vassia Atanassova, PhD
Bioinformatics and Mathematical Modelling Dept.
IBPhBME - BAS