

Johannes Schimunek

Institute for Machine Learning

Johannes Kepler University, Linz

Opponent Review “Protein solubility prediction using CNN and CNN-LSTM hybrid models” by Ekaterina Sysoykova

The thesis is concerned with machine learning techniques for protein solubility. The introduction section gives an overview about the term solubility in general, protein solubility and about prior machine learning related work. In the second section, the author presents the used model architectures, training and hyperparameter details and the data set. In the third section, a variety of metrics are introduced and, based on these metrics, the performances of the models are evaluated. A statistical analysis estimates the significance of the differences regarding one chosen metric.

The thesis treats a relevant problem.

Comprehensive references to relevant literature, a clear and extensive description of the touched fields and the usage of a variety of relevant metrics for performance evaluation build the strengths of this thesis. In the introduction part, the topic protein solubility is introduced well, referring to relevant literature. Here, also the importance of doing work in this direction is pointed out clearly. The author describes that a lot of machine-learning models are already in use in this field. Here, the different methods are named and cited carefully.

In contrast to many other publications in this field, not only one metric, but a variety of metrics are introduced and used to estimate the performance of the different implemented models.

There are some minor flaws in work.

There is a lack of information regarding the used data set. It is unclear how the test set is obtained. A RF is used as a baseline. Using RF for sequence data is not intuitive. Here the author should have provided more information, e.g. how the data is fed into the model.

In table 1, error bars are missing. The author performed a 10-fold cross-validation. Here it would have been easy, to compute the metrics from table one for every of these 10 runs.

In table 2, it is missing for which metric the statistics are provided. The related text passage includes this information though.

The author used Optuna for hyperparameter search, which is appreciated. The hyperparameter space and the chosen hyperparameters are given in the appendix. Exploring deep learning hyperparameters in only 25 iterations needs clarification.

Altogether, the positive aspects of this work clearly outweigh the flaws. Therefore, I recommend acceptance of this Bachelor's thesis.

Linz, 21-01-2022

Johannes Schimunek



Přírodovědecká
fakulta
Faculty
of Science

Jihočeská univerzita
v Českých Budějovicích
University of South Bohemia
in České Budějovice

OPPONENT'S REVIEW ON BACHELOR/DIPLOMA* THESIS

Name of the student: Ekaterina Sysoykova

Thesis title: Protein solubility prediction using CNN and CNN-LSTM hybrid models

Supervisor: Assist.-Prof. Mag. Dr. Günter Klambauer, Institut für Machine Learning, JKU
Philipp Seidl, MSc, , Institut für Machine Learning, JKU

Referee:

Referee's affiliation:

	Point scale ¹	Points
(1) FORMAL REQUIREMENTS		
Extent of the thesis (for bachelor theses min. 18 pages, for masters theses min. 25 pages), balanced length of the thesis parts (recommended length of the theoretical part is max. 1/3 of the total length), logical structure of the thesis	0-3	3
Quality of the theoretical part (review) (number and relevancy of the references, recency of the references)	0-3	3
Accuracy in citing of the references (presence of uncited sources, uniform style of the references, use of correct journal titles and abbreviations)	0-3	3
Graphic layout of the text and of the figures/tables	0-3	3
Quality of the annotation	0-3	3
Language and stylistics, complying with the valid terminology	0-3	3
Accuracy and completeness of figures/tables legends (clarity without reading the rest of the text, explanation of the symbols and labeling, indication of the units)	0-3	3
Formal requirements – points in total		21
(2) PRACTICAL REQUIREMENTS		
Clarity and fulfillment of the aims	0-3	2
Ability to understand the results, their interpretation, and clarity of the results, discussion, and conclusions	0-3	3
Discussion quality – interpretation of the results and their discussion with the literature (absence of discussion with the literature is not acceptable)	0-3	3
Logic in the course of the experimental work	0-3	3
Completeness of the description of the used techniques	0-3	2

* Choose one

¹ Mark as: 0-unsatisfactory, 1-satisfactory, 2-average, 3-excellent.

Experimental difficulty of the thesis, independence in experimental work	0-3	3
Quality of experimental data presentation	0-3	2
The use of up-to-date techniques	0-3	3
Contribution of the thesis to the knowledge in the field and possibility to publish the results (after eventual supplementary experiments)	0-3	3
Practical requirements – points in total		24
POINTS IN TOTAL (MAX/AWARDED)		48
		45 ²

Comments of the reviewer on the student and the thesis:

Suggestions and questions, to which the student has to answer during the defense.
Mistakes, which the students should avoid in the future:

Conclusion:

In conclusion, I

~~recommend / do not recommend~~*

the thesis for the defense and I suggest the grade 1.³

In Vienna; date Jan 24, 2020

signature

² Enter the number of points awarded.

³ You can suggest a grade, which can be modified during the defense based on the presentation. However, if the reviewer is not present at the defense, the grade will not be counted. Grades: excellent (1). Very good (2), Good (3), Unsatisfactory/failed (4).