ČESKÉ VYSOKÉ UČENÍ TECHNICKÉ V PRAZE

REVIEWER'S OPINION OF FINAL THESIS

I. IDENTIFICATION DATA

Thesis name:	**Anomaly detection in complex software architectures**
Author's name:	**Ondrej Borevec**
Type of thesis :	master
Faculty/Institute:	Faculty of Electrical Engineering (FEE)
Department:	Department of Computer Science and Engineering
Thesis reviewer:	Jan Brabec
Reviewer's department:	Department of Computer Science and Engineering

II. EVALUATION OF INDIVIDUAL CRITERIA

Assignment **challenging**

Evaluation of thesis difficulty of assignment.

I consider the assignment challenging as it requires a strong overlap of domain knowledge in software logging architectures and machine learning. Also, the thesis covers design of ML system end-to-end, from problem statement, through data collection to evaluation which takes into account the industrial constraints.

Satisfaction of assignment **fulfilled**

Assess that handed thesis meets assignment. Present points of assignment that fell short or were extended. Try to assess importance, impact or cause of each shortcoming.

All points in assignment were satisfied

Method of conception **correct**

Assess that student has chosen correct approach or solution methods.

The thesis includes strong study of related work and appropriate methods were chosen

Technical level **A - excellent.**

Assess level of thesis specialty, use of knowledge gained by study and by expert literature, use of sources and data gained by experience.

With a few minor objections the thesis is technically solid and deep. It skillfully combines the domain knowledge with machine learning and the approaches chosen are sound.

Objections:

1. On page 15 it is claimed that unsupervised learning has an unspoken requirement of balanced training dataset to be efficient. While I tend to agree with this in case of the default versions of unsupervised learning methods explored in the paper (PCA, k-means) I do not agree that unsupervised learning has this requirement in general. For such a claim, I would like to see an argument, either in the thesis or a citation that backs it up.
2. On page 18 the author refers to equation 2.4 which specifies the optimization problem for soft-margin SVM and claims that it would not be able to efficiently learn on linearly inseparable data and therefore kernel trick is used. The claim would probably be a valid shortcut if the equation specified hard-margin SVM but in case of soft-margin SVM it would be able to learn on linearly inseparable data. I understand that by 'efficiently learn' the author probably means 'good predictive performance' but I would expect more nuance in statements such as this.
3. On page 18 author calls the SVM sigmoid kernel 'neural network'. Why?
4. Figure 2.3 does not show performance of decision trees although section 2.2.1.4 implies that it does.
5. On page 29 it is claimed that optimizing accuracy would cause classifying all behavior as normal. This again, is an oversimplification and in limit case it is trivially not true (100 % accuracy requires all samples to be classified correctly). Accuracy is not a good measure for evaluation on imbalanced problems but lots of state-of-the-art methods suitable for imbalanced classification optimize accuracy during training phase.
6. Evaluation should include ROC curves which are not affected by the dataset imbalance to help better estimate the expected performance of the classifier beyond this particular test dataset.

Formal and language level, scope of thesis **C - good.**

Assess correctness of usage of formal notation. Assess typographical and language arrangement of thesis.

The thesis is written mostly in good English but the final product seems a bit rushed and could use additional round of corrections. There is a not a small number of typos. For example: pg.7 crows -> crowd, pg.8 servers -> serves, is be deterministic -> is deterministic, cross -> across, pg. 10 afford -> effort, …

Also, there are several phrases that are not appropriate in formal academic writing.

Chapter: Motivation

1. You can definitely image (sic), we were facing some problems nearly every day […]
2. [..] since nobody in the company could monitor their own services (hyperbole is not appropriate in formal writing)

Page 15

1. We would like to give a quick shout out to all open-source monitoring software [...]

Page 43

1. Therefore we decided to work with this model type rather than with a *fancy* neural network model.

Selection of sources, citation correctness **A - excellent.**

Present your opinion to student's activity when obtaining and using study materials for thesis creation. Characterize selection of sources. Assess that student used all relevant sources. Verify that all used elements are correctly distinguished from own results and thoughts. Assess that citation ethics has not been breached and that all bibliographic citations are complete and in accordance with citation convention and standards.

The thesis cites 64 sources and also includes quite a lot of relevant links to mentioned software packages in footnotes. The sources are relevant.

Additional commentary and evaluation

Present your opinion to achieved primary goals of thesis, e.g. level of theoretical results, level and functionality of technical or software conception, publication performance, experimental dexterity etc.

III. OVERALL EVALUATION, QUESTIONS FOR DEFENSE, CLASSIFICATION SUGGESTION

Summarize thesis aspects that swayed your final evaluation. Please present apt questions which student should answer during defense.

The thesis is technically sound and I appreciate that it tackles a real problem that is present in the industry and thus has immediate applications. A major contribution is the creation and open-sourcing of the dataset. I also appreciate that the usefulness and the experience of the users of the system is considered (for example, throttling of events to 1 per 5-minute window). On the other hand, the text could probably use another round of revisions to correct the style.

- The emphasis in the thesis is on avoiding false-negatives but what is the cost of a false positive alarm? On the evaluation dataset the methods have around 20 % precision, in a real setting with different

imbalance-ratio the precision can be much lower. Is this precision acceptable for the system to be applied in the wild?

- In Chapter 3, the problem is formalized as a Neyman-Pearson task, but I have trouble finding how this formalization is used by the proposed algorithm. How does it optimize the Neyman-Pearson problem?

I evaluate handed thesis with classification grade **B - very good.**

Date: **14.6.2019** Signature: