THESIS REVIEWER'S REPORT

I. IDENTIFICATION DATA

Thesis title:	**Anomaly Detection Methods for Log Files**
Author's name:	**Martin Koryťák**
Type of thesis :	master
Faculty/Institute:	Faculty of Electrical Engineering (FEE)
Department:	CS
Thesis reviewer:	Gustav Šír
Reviewer's department:	CS

II. EVALUATION OF INDIVIDUAL CRITERIA

Assignment **ordinarily challenging**

How demanding was the assigned project?

It seems like a fairly regular ML assignment to me – the task, data and models have all been generally known beforehand.

Fulfilment of assignment **fulfilled**

How well does the thesis fulfil the assigned task? Have the primary goals been achieved? Which assigned tasks have been incompletely covered, and which parts of the thesis are overextended? Justify your answer.

I see absolutely no problem here, only HDFS log files are experimented with, but that's completely ok.

Methodology **correct**

Comment on the correctness of the approach and/or the solution methods.

Very solid approach trying out a range of modern models with a sound methodology and project structure.

Technical level **B - very good.**

Is the thesis technically sound? How well did the student employ expertise in the field of his/her field of study? Does the student explain clearly what he/she has done?

It is very nice in trying out some of the latest DL models and their combinations. I have some small doubts about the student's clear understanding behind their inner workings and the respective (supervised) ML methodology, as detailed later in comments. Nevertheless these are only minor issues.

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

Are formalisms and notations used properly? Is the thesis organized in a logical way? Is the thesis sufficiently extensive? Is the thesis well-presented? Is the language clear and understandable? Is the English satisfactory?

Very nicely structured, very clear description of the experimental protocol. Only in very few cases the flow of the description confused me (e.g. describing CNNs in a respectively named section, then RNNs in next sentence, then back to CNNs. Similarly in few other places). Likewise the choice of some unusual English words (probably picked via direct CZ-EN translation) and phrases got me sometimes confused, but the English level is general really good and polished.

Selection of sources, citation correctness **B - very good.**

Does the thesis make adequate reference to earlier work on the topic? Was the selection of sources adequate? Is the student's original work clearly distinguished from earlier work in the field? Do the bibliographic citations meet the standards?

I was expecting sources in few places (e.g. the pictures), and some use of sources seemed to misinterpret the original a bit (see comments later), but it's generally fine, following the citation etiquette standards.

Additional commentary and evaluation (optional)

Comment on the overall quality of the thesis, its novelty and its impact on the field, its strengths and weaknesses, the utility of the solution that is presented, the theoretical/formal level, the student's skillfulness, etc.

This is a very nice, practically oriented, thesis. The structure is well-thought-out, only reflecting the well-structured project itself. I'm not quite sure how it compares to state-of-the-art, but it generally seems at such a level to me, hence the grading. Nevertheless, I have some comments/questions:

„[Receptive fields + Pooling] …which leads to the reduction of trainable parameters [in CNNs]"

- what really reduces the parameter count in CNNs is the *weight sharing* induced by the repeated application of the *same* convolutional filter (w.r.t MLPs of an equivalent structure)

„fully connected layers at the end of the network since they exponentially increase the complexity of a particular architecture„

- I do not think there is anything exponential about matrix multiplication?

„[L2 regularization] penalizes „jagged" weight vectors [7]"

- that's a bit misleading, it simply penalizes large values, there is no notion of their structure
- I do not think they say that in [7], do they?

Similarly: *"Semi-supervised methods assume that all training data points belong to the normal class.[27]"*

Whereas in [27]: **"**…assume that the training data has *labeled instances* for only the normal class"

- which is different

Consequently, I got confused a lot later in the thesis – your HDFS dataset is highly imbalanced, but completely labeled: *"each collection of logs is annotated by the normal or the anomalous label"* yet you keep talking about semi-supervised task (I get that you use unsupervised model training, but the task is supervised)

Consequently: *Surprisingly, we assume that the number of epochs is also a hyperparameter*

- very confusing!

…This is caused by the nature of semisupervised machine learning as there is no validation data set

- But there is: *„The training and validation data sets follow the exact split ratio (10 : 1)"*

Ultimately leading to how you turn the unsupervised training to supervised labeling through finding the optimal decision threshold (theta) – naturally, this is a *parameter* and the optimal theta-*search* should be part of the *training*! (not validation - it is not a hyperparameter – see your own definition of it!). Consequently, it's a normal supervised training task, and you can simply directly train w.r.t. the F1 score maximization. Moreover, you (probably only mildly) overfit the validation data by doing this.

Ad-hoc: „*The Skip-gram model learns to predict a word embedding vector using the surrounding neighborhood. On the contrary, the CBOW model predicts a word embedding vector using its context […consisting of the word's fixed size neighborhood]"*

- These are 2 *equivalent* formulations (i.e. that's not how the CBOW works, it works in reverse)
- Your use of the „On the contrary" phrase got me confused in few other places, too.

III. OVERALL EVALUATION, QUESTIONS FOR THE PRESENTATION AND DEFENSE OF THE THESIS, SUGGESTED GRADE

Summarize your opinion on the thesis and explain your final grading. Pose questions that should be answered during the presentation and defense of the student's work.

Provide clarifications on the comments above. Then the grade that I award for the thesis is **A - excellent.**

Date: **28.5.2021** Signature: