CTU
CZECH TECHNICAL UNIVERSITY IN PRAGUE

THESIS SUPERVISOR'S REPORT

I. IDENTIFICATION DATA

Thesis title:	**Analyzing the execution of malware in a sandbox using hierarchical multiple instance learning**
Author's name:	**Dominik Kouba**
Type of thesis :	Master Thesis
Faculty/Institute:	Faculty of Electrical Engineering
Department:	Department of Computer Science
Thesis reviewer:	Tomáš Pevný
Reviewer's department:	Department of Computer Science

II. EVALUATION OF INDIVIDUAL CRITERIA

Assignment **A**

How demanding was the assigned project?

The assignment was very demanding, since the student has to solve at least two very different problems: the configuration of an environment for capturing execution of malware in the sandbox and creation of classifiers on structured data and explaining their decisions.

Fulfilment of assignment **A**

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.

The assignment was fulfilled completely.

Activity and independence when creating final thesis **A**

Assess whether the student had a positive approach, whether the time limits were met, whether the conception was regularly consulted and whether the student was well prepared for the consultations. Assess the student's ability to work independently.

The student was very active. For discussions of his progress, which has been held every week, he has come with a list of problems (and questions) he needs to discuss.

Technical level **A**

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

The level of the thesis is good. The student has demonstrated he can master theoretical and also the practical aspect of the problem.

Formal level and language level, scope of thesis **C**

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?

The language and organisation of the thesis is the weakest part of the work. The text would benefit from a better English formulations and also better structure, particularly from rethinking what is needed for the story and what can be omitted. But in defence of the student, he has tried to describe in a single coherent text two different topics (see above), which is everything but trivial.

Selection of sources, citation correctness **A**

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?

110 citations correlates with the volume of information the student went through and also with his meticulous not to omit any citation.

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.

Please insert your comments here.

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

As I have mentioned on the beginning, the scope of the thesis is enormous, as just a configuration of a distributed environment for capturing logs of malware execution is everything but trivial and can be a thesis of its own. The student has used these captures to evaluate, how a particular behaviour of malware can be detected from a portion of logs, which has not been used to create the labels. This is very interesting and important, as in some cases the information to create labels (API calls) might not be available, and being able to detect from different sources of data increases the applicability. The student also used a relatively new (at the time of writing closed sourced) tool for explaining hierarchical data to validate, how the sources of detection correlates with the true causes of the behaviour. In his analysis he points to the limit of the statistical analysis, specifically that it is difficult to differentiate between correlation and causality. Although I have pointed out that the writing aspect of the thesis can be improved, I think it should not have an effect on the final evaluation.

The grade that I award for the thesis is **A**

Date: 31.5.2021 Signature: