ČVUT
ČESKÉ VYSOKÉ
UČENÍ TECHNICKÉ
V PRAZE

REVIEWER‘S OPINION OF FINAL THESIS

I. IDENTIFICATION DATA

Thesis name:	**Analysing the execution of malware in a sandbox using hierarchical multiple instance learning**
Author’s name:	**Dominik Kouba**
Type of thesis :	master
Faculty/Institute:	Faculty of Electrical Engineering (FEE)
Department:	Department of Computer Science
Thesis reviewer:	Professor Lorenzo Cavallaro
Reviewer’s department:	King's College London

II. EVALUATION OF INDIVIDUAL CRITERIA

Assignment **extraordinarily challenging**

Evaluation of thesis difficulty of assignment.

This is an extraordinarily challenging project. Not only the work clearly requires to dive into open research questions, but it also requires to cross the boundaries of different – often time disconnected – disciplines. The use of machine learning in cybersecurity it is nowadays a well-accepted fact. Both industry and academia has been building success stories on this. However, a proper – and successful – project in this domain should be supported by a clear evidence of solid understanding of these two disciplines on their own as well as when considered together.

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.

Following up from previous criteria, Domink shows a solid understanding of cybersecurity as well as machine learning concepts. The project represents a fairly complex yet mature pipeline to enable theoretical reasoning as well as practical (i.e., impactful) considerations. Of particular interest is the consideration of interpretability in addition to traditional metrics (of performance).

Method of conception **outstanding**

Assess that student has chosen correct approach or solution methods.

The method is well-motivated and the report is clearly written. It has been a real joy to read the project report. The way that Dominik introduces and motivate the methodology is sound and pretty insightful. There are several ways in which one would follow-up this work, e.g., exploring concept drift and adversarial attacks more in details, but this project provides a solid baseline to compare against.

Technical level **B - very good.**

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

The work requires solid knowledge of maths as well as systems. As far as I can tell, Dominik was able to deliver on both with clear precision and outstanding results.

Formal and language level, scope of thesis **A - excellent.**

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

To the best of my knowledge, formalism is in line with the formalism the ML and security community rely on. The two communities may use slightly different terminology at time but one can understand the correct meaning by considering the context (e.g., whether more ML or cybersecurity-centric).

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

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 report presents the right level of sources and citation. Perhaps, there is some room one may want to consider wrt open problems in the context of Trustworthy ML (e.g., robustness against concept drift, adversarial ML) that could have been incorporated as citations properly contextualized in future work/discussion section.

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.

This is truly a remarkable thesis. There’s a clear motivation and intuition that guides the reader towards accepting the proposed solution. The work is solid and shows promising results. The problem space has several unanswered questions and open problems and the work represents an interesting perspective to explores promising approaches in ML to apply in malware classification problems. The thesis starts exploring some properties of Trustworthy ML (i.e., focus on model interpretability), which paves the way towards principled reasoning in this space.

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.

I evaluate handed thesis with classification grade **A - excellent.**

This is excellent work at the level of MSc thesis, as outlined in the supporting statements above to the evaluation criteria. Some questions that it would be interesting Dominik to answer may be (almost all of them randomly) picked by the following list:

- HMill seems to fall in context of representation learning. Wouldn’t a graph embedding (e.g., inductive GNN) be useful to identify stronger (control and/or data-flow) relationships than those expressed in a JSON (behavioral) report?
- Table 7.2 seems to contradict Table 7.1 in terms of class ratio. Table 7.1 report 1:1, whereas Table 7.2 different ratios (P). It would be great if this could be clarified (in the text too).
- Following up on the previous point, is there a risk of experimental bias as outlined in [1]? For instance, are all the samples (goodware and malware) drawn from the same timeframe? Is the testing class ratio manipulated artificially? (There’s a risk of inflating results otherwise.)
- What would be needed to perform a time-aware evaluation as outlined in [1] and to assess to what extent HMill is robust against concept drift [2, 3, 4]?

[1] Pendlebury et al. TESSERACT: Eliminating Experimental Bias in Malware Classification across Time and Space. USENIX Sec 2019.

[2] Jordaney et al. Transcend: Detecting Concept Drift in Malware Classification Models. USENIX Sec 2017

[3] Barbero et al. Transcending Transcend: Revisiting Malware Classification in the Presence of Concept Drift. arXiv 2020

[4] Deo et al. Prescience: Probabilistic Guidance on the Retraining Conundrum for Malware Detection. AISec 2016

Date: **7.6.2021**

Signature: Lorenzo Cavallaro