FACULTY
OF INFORMATION
TECHNOLOGY
CTU IN PRAGUE

# Review report of a final thesis

| | |
|---|---|
| **Reviewer:** | Mgr. Tomáš Rabas |
| **Student:** | Mehmet Efe Zorlutuna |
| **Thesis title:** | Unsupersived Instance Selection for Malware Detection |
| **Branch / specialization:** | Computer Security 2021 |
| **Created on:** | 5 June 2023 |

## Evaluation criteria

### 1. Fulfillment of the assignment

▸ [1] **assignment fulfilled**
[2] assignment fulfilled with minor objections
[3] assignment fulfilled with major objections
[4] assignment not fulfilled

"Instructions:"
"1) Study the state-of-the-art unsupervised instance selection algorithms."
- a list of some state-of-the-art algorithms is written in chapter 1.3.6.
Chapters 1 and 4 provide a good knowledge base not only for unsupervised machine learning algorithms but more generally for AI and malware detection.
"2) Try to propose new or modify existing unsupervised instance selection algorithms."
- In Chapter 2 is described a new instance-detection algorithm called Nearest Cluster Enemy (NCE). Its more thorough description with implementation details is in Chapter 3.
"3) Use existing libraries or implement at least two unsupervised instance selection algorithms for malware detection."
- Fulfilled using python library sklearn
"4) Compare and discuss the experimental results in terms of the reduction rate, the accuracy, and the computational time."
- Results are presented in Chapter 6 and some details about testing in Chapter 5.

### 2. Main written part 87 /100 (B)

Intermediate shortcoming:
- In Chapter 1 - Background Information (only 5 citations), and Chapter 4 - Test Background Information (only 4 citations) there would be good to add much more citations. For example, from chapter 1.1.1 until 1.3.5 there are not any citations which I see as a little problematic if one would want to follow that resource to gain more information about that and also to assess the originality of the work.

Small shortcomings or suggestions for improvements:
- During a description of algorithms in Chapter 1.3.6 (State-of-Art Instance Selection Algorithms) it would be good to state which categories (defined in chapters 1.3.4 and 1.3.5) they belong to. I.e. state if they are supervised/unsupervised and filtering/wrapper method.
- In Chapter 2, descriptions of axes x and y would be helpful. Preferably in a figure itself, or at least in the text when the figure is mentioned.
- In Chapter 2.1 - Preprocessing and Feature Extraction, a graph after preprocessing step (visualization of the second step) is not included - if it does not change from previous steps, it is worth mentioning it.
- In Chapter 2.4 - Elimination, two (original ?) elimination techniques are mentioned. Although the core idea/justification/heuristics behind each elimination technique is not provided. In other words, an explanation to the questions "why such a technique or method should work?" and "how the author came up with this idea?" should be provided.
- In Chapter 2.4 it is referred to the threshold, even though it is not described there, but a few chapters later. Also, it would be helpful to know what threshold was used for the example.

- In Chapter 3.2.3 - Clustering, the second paragraph consists of only one long sentence. It is not grammatically correct and very difficult to understand:
"The algorithm works as, clustering starts with a random instance, every instance in the
epsilon range included in the group to create the cluster, then this is repeated for each new
group member recursively when there is no instance left in the epsilon range, group members
counted if the count is smaller than the minimum sample number, corresponding instances are
considered noise if the count is equal to or greater than the minimum sample number, a new
cluster is formed."
- In Chapter 3.3.2 - Elimination technique 2 is a threshold chosen to be "the minimum sample number described in the clustering step". An explanation of why would be helpful also with a short discussion on other choices.
- In Chapter 5.2 - Training and Testing Data Sets as in Chapter 6.1.4 - Execution Time, when discussing computation time, I would suggest using CPU time and characterising the device's computational power and other characteristics for a more precise view of computational complexity.

## 3. Non-written part, attachments 87 /100 (B)

## 4. Evaluation of results, publication outputs and awards 90 /100 (A)

# The overall evaluation 88 /100 (B)

The final thesis fully satisfies the assignment.
The student suggested her own method and compared it to another.
The thesis does not tend to exaggerate the advantages of a proposed method but rationally evaluates its properties.

It provides all the necessary information to understand the results and gain insight into the field for the reader.
It is easy to read and understand, even quite captivating.
It is clear that the student fully understands the content.
Occasionally, some thoughts are repeated too many times so it starts to be tedious (even though to some extent it is necessary for good understandability).
Similarly, testing results could be presented in a shorter manner, using graphs and more-dimensional tables.
The originality and heuristics/ideas/justifications behind the proposed method could be explained better (although the depth of an explanation seems to be appropriate for the field).

## Questions for the defense

1) What sources did a student use during writing Chapter 1 - Background Information (only 5 citations), and Chapter 4 - Test Background Information (only 4 citations), and why did the student mostly not cite them?
2) Is the idea behind the proposed two elimination techniques an original work of the student or a student was (partly) inspired somewhere else - where?

# Instructions

## Fulfillment of the assignment

Assess whether the submitted FT defines the objectives sufficiently and in line with the assignment; whether the objectives are formulated correctly and fulfilled sufficiently. In the comment, specify the points of the assignment that have not been met, assess the severity, impact, and, if appropriate, also the cause of the deficiencies. If the assignment differs substantially from the standards for the FT or if the student has developed the FT beyond the assignment, describe the way it got reflected on the quality of the assignment's fulfilment and the way it affected your final evaluation.

## Main written part

Evaluate whether the extent of the FT is adequate to its content and scope: are all the parts of the FT contentful and necessary? Next, consider whether the submitted FT is actually correct – are there factual errors or inaccuracies?

Evaluate the logical structure of the FT, the thematic flow between chapters and whether the text is comprehensible to the reader. Assess whether the formal notations in the FT are used correctly. Assess the typographic and language aspects of the FT, follow the Dean's Directive No. 52/2021, Art. 3.

Evaluate whether the relevant sources are properly used, quoted and cited. Verify that all quotes are properly distinguished from the results achieved in the FT, thus, that the citation ethics has not been violated and that the citations are complete and in accordance with citation practices and standards. Finally, evaluate whether the software and other copyrighted works have been used in accordance with their license terms.

## Non-written part, attachments

Depending on the nature of the FT, comment on the non-written part of the thesis. For example: SW work – the overall quality of the program. Is the technology used (from the development to deployment) suitable and adequate? HW – functional sample. Evaluate the technology and tools used. Research and experimental work – repeatability of the experiment.

## Evaluation of results, publication outputs and awards

Depending on the nature of the thesis, estimate whether the thesis results could be deployed in practice; alternatively, evaluate whether the results of the FT extend the already published/known results or whether they bring in completely new findings.

## The overall evaluation

Summarize which of the aspects of the FT affected your grading process the most. The overall grade does not need to be an arithmetic mean (or other value) calculated from the evaluation in the previous criteria. Generally, a well-fulfilled assignment is assessed by grade A.