THESIS REVIEWER'S REPORT

I. IDENTIFICATION DATA

Thesis title:	Identifying Suspicious Behavior of Network Devices Using Machine Learning Methods
Author's name:	**Zubaryk Katsiaryna**
Type of thesis:	bachelor
Faculty/Institute:	Faculty of Electrical Engineering (FEE)
Department:	DEPARTMENT OF TELECOMMUNICATION ENGINEERING
Thesis reviewer:	Ing. Cenek Skarda
Reviewer's department:	Cisco

II. EVALUATION OF INDIVIDUAL CRITERIA

Assignment **challenging**

How demanding was the assigned project?

The assigned project was conceived as an applied ML research task. Considering the student background, the project was challenging, requiring a substantial effort and an extensive study of ML algorithms and techniques.

Fulfilment of assignment **fulfilled with major objections**

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.

Completion of the thesis should be divided into two parts - what the student has done and the quality of submitted thesis. The author successfully walks through the entire ML-based solution proposing unsupervised anomaly detection on a real categorical data. That includes initial data analyses, design of feature space encoding, selection of several clustering algorithms and metrics and the estimation of anomalous devices based on the distance from cluster centroids.

Unfortunately, the student heavily underestimated the time and effort needed for writing the first technical text. Without the discussion with the consultant of the thesis, I could not tell which of the above-mentioned tasks were actually covered and with what results.

Methodology **correct**

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

The chosen approach and solution methods are correct. Unfortunately, the solution is not summarized anywhere in the thesis, making it hard to follow.

Technical level **F - failed.**

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?

The text of the work usually does not sound technical. The thesis is quite long, but this is mainly due to the excessive number of figures, tables and code samples that are not well described, incorrectly formatted, without proper captions, etc. throughout the thesis. The actual text is about 17 pages long. It is extremely difficult to establish what the student was doing and how the tasks relate to each other. In most cases, the text is a mixture of references, code descriptions, and results. The results are not well described. Several times they end with a sentence like "Which makes my hypothesis correct." without mentioning the hypothesis beforehand and without arguing why it should be true (which it probably isn't in many cases).

Formal and language level, scope of thesis **F - failed.**

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 passive, plural and singular alternate in the text. The relatively short text refers heavily to figures, code samples and tables without properly describing what they refer to. The captions, which are short, misleading, or just plain wrong, do not help (e.g., Figures A.5-A.8 state "polar graph" when they are histograms, with no description of the axes). A huge proportion of the figures is questionable - "what story do they build?", "are they used later for discussion?", "do they help the reader understand the text and concepts?". In addition, the figures often exceed the page boundaries or are illegibly small.

Chapters contain unrelated concepts, paragraphs do not build on each other, and in many cases, sentences are not consistent. Thesis contains many inaccurate or misleading statements. It is obvious which parts were written in advance and which parts (around 80% of the thesis) were written in haste.

Selection of sources, citation correctness **C - 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?

Citations are correct across the whole thesis. Selection of sources is partially relevant, partially irrelevant, e.g., “unsupervised identification” isn’t commonly used term (unsupervised learning is), so the cited papers (e.g., [2]) aren’t from the field of the thesis and so, it is partially misleading.

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.

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.

The thesis has the potential to become a great bachelor thesis. The used ML algorithms are well known, but their applicability in the field, in which they have been applied by the author, is not yet well explored. While the topic and results are very interesting, the thesis summarizing the work is in a terrible state. The actual text is short, stuffed with misleading, poorly formatted images. The thesis contains errors and inaccuracies, does not describe the results successfully, and does not move from high concepts to details. Only a few sections follow the principles of a technical text.

Question A: The work investigated multiple algorithms, their hyperparameters and metrics. Which configuration seems most promising and why? Can you provide a comparison of the parameters on the final set of devices identified as anomalous?

Question B: Did you test other than one-hot encoding? E.g., by using number of realizations of individual events?

Question C: What is the reason to pick two out of three metrics for k-means and hierarchical clustering (different subset for each – section 4.3.1)?

Question D: What was the goal of dimensionality reduction? Is it the efficacy improvement or the limitation of used unsupervised methods?

Question E: The explainability of unsupervised methods is highly valuable. Can you show dendrograms for hierarchical clustering? Did you test t-SNE or some similar method for visualization of multidimensional data? Polar plot idea is great – how they perform on visualization of multiple clusters?

The grade that I award for the thesis is **F - failed.**

Date: **7.6.2023**

Signature: Ing. Cenek Skarda