OULUN YLIOPISTO
UNIVERSITY of OULU

Explainable AI for Log-Based Anomaly Detection in Security Monitoring: Reasoning Pipelines and Cross-Dataset Evaluation.

University Of Oulu
Faculty of Information Technology and
Electrical Engineering
Master’s Thesis

Emmanuel Ikwunna
9th November 2025

Abstract

Abstract is needed to sum the master's thesis up. The abstract is to be uploaded into Optima before the final grading of the thesis. Please find the current information about the format given in Optima.

The guide includes instructions for students. It is written keeping in mind the idea that the user may utilise it e.g. by pasting his or her text on the current text. The contents include information about formatting the text, positioning tables and figures, among other things. In addition, the use of proper literature is instructed. Even if there is no strict structure for the thesis, a recommendation is offered in this guideline.

One important guideline for the text is that do not write too short paragraphs. For instance, if there is only one sentence in a paragraph, the sentence must be really important and influential to form a paragraph of its own.

It is not possible to provide information in a guideline like this for all issues related to master's thesis. For example, the research process, ways to acquire research material and its analysis are excluded in the guideline. On the other hand, a structure for a research plan is provided in the appendices.

Keywords
first keyword, second keyword, other keywords

Supervisor
Title, position First name Last name

Foreword

The foreword is not instructed by the supervisors. In other words, the student may write in this section what she or he wants to share with readers. However, it is a custom to thank all those who have contributed to the research somehow. When acknowledging people, their affiliations are given (e.g. Professor, University Lecturer, Adjunct Professor, Mrs.) This guideline is based on the previous version that was written in Finnish and finalised by Dr. Lasse Harjumaa in January 2007. This version is to replace the earlier version. I want to thank all those people who have contributed to the earlier versions and this newest version, the first written in English. Hopefully this guideline will serve both students and faculty with its instructions that include both formal and informal regulations and recommendations. In the first phase, the constructive comments are received with pleasure by raija.halonen@oulu.fi. Oulu, January 10, 2011

Raija Halonen Oulu, March 10, 2020

Contents

1 Introduction

In the thesis we follow the style introduced by The American Psychological Association (APA). The APA style can be found easily in the Internet and some sites provide a quick guide, too. E.g. `http://www.waikato.ac.nz/library/learning/g_apaguide.shtml` and `http://owl.english.purdue.edu/owl/resource/560/01/` are useful links.

It is important to follow given instructions. In academic theses, not only the content but also the format is important. Generally every academic publication forum requires that the publications follow their guidelines. In the theses accepted in the Department of Information Processing Science the format is APA. Currently there are several editions published from APA. The general rule is that the latest available edition is applied. Currently the newest edition is 6th. If a thesis is already in process it is not needed to transfer it into a newer edition of APA. Whichever you apply, do it consistently.

In addition to teach the students to follow given formal instructions, the guideline aims to unify and standardise the outlook of the theses made in the department. The guideline also enables the supervisors to focus on the content of the theses as the students already consider the outlook and format themselves. In this sense, it is a question of available resources for supervision and guidance.

The use of language and grammar cannot be discussed in detail in this kind of guide. However, the writing style should meet the general academic writing styles in the sense that no causeries are accepted or other lightweight texts such as jokes or rumblings. In other words, in academic theses all writing must be appropriate and reasonable. There are several guidebooks for academic writing available in the Oulu University Library, for example, and in the Internet. For those who write their thesis in Finnish there are books such as Tieteellinen kirjoittaminen. The style reference by APA (American Psychological Association, 2010) offers fruitful practical hints for writing thesis in English.

As the guideline is written according to the instructions, it enables the students to copy their text (without format) on the document and thus get their text into the right format. The format is to be used in the Bachelor's Theses and in the Master's Theses. In case of other theses, essays or reports it is recommended that the students inquire their teachers if the guideline is to be followed or not.

The structure of the guideline is as follows. The formal instructions for different topics are presented next. This is followed by examples of references and their use. After that the structure of theses and its writing style is discussed briefly. The guideline ends with a summary.

2 BACKGROUND

2.1 Log Anomaly Detection

System log files, which are ubiquitously generated by networked computer systems, contain valuable information that is essential for monitoring system stability and security [1, 2]. Logs describe detailed system events at runtime and chronologically record the intentions of users [3]. As systems become increasingly complex, these logs serve as a crucial resource for fault diagnosis, performance monitoring, security auditing, and transaction tracing [4].

The primary purpose of log anomaly detection is to protect digital infrastructures by identifying abnormal activities, such as network intrusions, from the enormous volumes of event logs. Anomalies can be defined as patterns in data that significantly deviate from the expected behavior of the system. Detecting these deviations is crucial for maintaining system reliability and preventing severe disruptions or financial losses, as global cybercrime costs are estimated to reach trillions of euros annually [5].

Log analysis, and consequently anomaly detection, faces several significant challenges primarily driven by the nature and scale of the data:

- **Volume:** System logs are large-scale data collected in real-time [6]. The sheer volume of logs has grown rapidly, often reaching 50 GB (120–200 million lines) per hour for large-scale services, making manual inspection and traditional processing infeasible [3].

- **Variety and Complexity:** Logs are typically unstructured or semi-structured text files generated by logging statements in source code [2]. Because developers are allowed to write free-text messages, the format and semantics of logs vary significantly from system to system, leading to high-dimensional features with complex interrelationships. This complexity and diversity increase the difficulty of accurate anomaly detection [7].

- **Velocity (Timeliness):** For anomaly detection to be useful, it must be timely, requiring decisions to be made in a streaming fashion to allow users to intervene in ongoing attacks or performance issues. Offline methods that require multiple passes over the entire log data are thus unsuitable for real-time security monitoring [2].

Due to the challenges of volume and complexity, the adoption of automated log analysis has become imperative to efficiently process and interpret vast corpora of logs.

2.2 Traditional Methods

Early log anomaly detection efforts relied heavily on human expertise [4]. As the volume of logs grew, research shifted toward automated, data-driven methods, broadly categorized into rule-based systems and statistical approaches, many of which depend on logs first being converted into a structured format through a process known as log parsing [7].

Log parsing is a critical precursor step where raw, unstructured log messages are transformed into structured data, typically by extracting a constant part, called the log template (or log key), and identifying the variable parts (parameters). The parser Spell, for example, is an online streaming parser that utilizes the Longest Common Subsequence (LCS) technique to dynamically identify and update log patterns. Tools like DeepLog rely on log parsing methods like Spell to generate log templates for their inputs [2, 8].

2.2.1 Rule-Based Systems (Regex, Signatures)

Rule-based methodologies were among the first attempts to automate log analysis to reduce human error. These methods typically rely on explicitly defined rules, patterns, or known indicators of abnormal behavior, often requiring specific domain knowledge from human experts.

- **Keyword Matching and Regular Expressions (Regex):** Early rule-based systems focused on matching specific keywords (e.g., "error," "failed") or using regular expressions to flag anomalous log entries [9]. However, relying solely on keywords or structural features often prevents a large portion of log anomalies from being detected and can lead to unnecessary alarms (alarm fatigue) if the system constantly evolves [9, 10]. Furthermore, manually designing and maintaining regular expressions is prohibitive given the rapid increase in log volume and frequent system updates [7].

- **Invariant Mining (IM):** Invariant mining is another traditional approach that captures co-occurrence patterns between different log keys [2, 10]. This method defines a window (time or session based) and detects whether certain mined quantitative relationships, or invariants, hold true within that window (e.g., ensuring that the count of "file open" logs equals the count of "file close" logs in a normal condition) [10]. IM is typically characterized as an unsupervised offline method [2].

2.2.2 Statistical Methods (Clustering, PCA)

Statistical methods leverage mathematical principles to identify normal patterns from data volumes, and flag deviations statistically likely to be anomalous. These methods generally operate on a generated numeric vector representation of the logs, often discarding parameter values and only using log keys and their counts [2]. Most statistical methods rely on initial log parsing, where raw log messages are converted into structured, numeric representations, such as event count vectors [11].

- **Principal Component Analysis (PCA):** PCA is a widely used statistical method for log anomaly detection. PCA is a linear transformation technique used to convert a set of correlated variables into a set of uncorrelated variables, known as principal components [6, 12].

-

References

[1] Crispin Almodovar, Fariza Sabrina, Sarvnaz Karimi and Salahuddin Azad. ‘Can Language Models Help in System Security? Investigating Log Anomaly Detection using BERT’. In: *Proceedings of the 20th Annual Workshop of the Australasian Language Technology Association*. Ed. by Pradeesh Parameswaran, Jennifer Biggs and David Powers. Adelaide, Australia: Australasian Language Technology Association, Dec. 2022, pp. 139–147. URL: https://aclanthology.org/2022.alta-1.19/.

[2] Min Du, Feifei Li, Guineng Zheng and Vivek Srikumar. ‘DeepLog: Anomaly Detection and Diagnosis from System Logs through Deep Learning’. In: *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*. CCS ’17. Dallas, Texas, USA: Association for Computing Machinery, 2017, pp. 1285–1298. ISBN: 9781450349468. DOI: 10.1145/3133956.3134015. URL: https://doi.org/10.1145/3133956.3134015.

[3] Hongcheng Guo, Jian Yang, Jiaheng Liu, Jiaqi Bai, Boyang Wang, Zhoujun Li, Tieqiao Zheng, Bo Zhang, Junran peng and Qi Tian. *LogFormer: A Pre-train and Tuning Pipeline for Log Anomaly Detection*. 2024. arXiv: 2401.04749 [cs.LG]. URL: https://arxiv.org/abs/2401.04749.

[4] Hong Huang, Wengang Luo, Yunfei Wang, Yinghang Zhou and Weitao Huang. ‘LogCTBL: a hybrid deep learning model for log-based anomaly detection’. In: *The Journal of Supercomputing* 81 (Jan. 2025). DOI: 10.1007/s11227-025-06926-3.

[5] European Parliament. *Cybercrime in the EU: Threats, Trends and Policy Responses*. Tech. rep. According to an EU briefing, the annual global cost of cybercrime was estimated at approximately €5.5 trillion in recent years. Accessed November 9, 2025. European Parliamentary Research Service, 2024. URL: https://www.europarl.europa.eu/RegData/etudes/BRIE/2024/760356/EPRS_BRI(2024)760356_EN.pdf.

[6] Yukyung Lee, Jina Kim and Pilsung Kang. ‘LAnoBERT: System log anomaly detection based on BERT masked language model’. In: *Applied Soft Computing* 146 (2023), p. 110689. ISSN: 1568-4946. DOI: https://doi.org/10.1016/j.asoc.2023.110689. URL: https://www.sciencedirect.com/science/article/pii/S156849462300707X.

[7] Pinjia He, Jieming Zhu, Zibin Zheng and Michael R. Lyu. ‘Drain: An Online Log Parsing Approach with Fixed Depth Tree’. In: *2017 IEEE International Conference on Web Services (ICWS)*. 2017, pp. 33–40. DOI: 10.1109/ICWS.2017.13.

[8] Min Du and Feifei Li. ‘Spell: Streaming Parsing of System Event Logs’. In: *2016 IEEE 16th International Conference on Data Mining (ICDM)*. 2016, pp. 859–864. DOI: 10.1109/ICDM.2016.0103.

[9] Harold Ott, Jasmin Bogatinovski, Alexander Acker, Sasho Nedelkoski and Odej Kao. *Robust and Transferable Anomaly Detection in Log Data using Pre-Trained Language Models*. 2021. arXiv: 2102.11570 [cs.AI]. URL: https://arxiv.org/abs/2102.11570.

[10] Weibin Meng, Ying Liu, Yichen Zhu, Shenglin Zhang, Dan Pei, Yuqing Liu, Yihao Chen, Ruizhi Zhang, Shimin Tao, Pei Sun and Rong Zhou. 'LogAnomaly: Unsupervised Detection of Sequential and Quantitative Anomalies in Unstructured Logs'. In: *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*. International Joint Conferences on Artificial Intelligence Organization, July 2019, pp. 4739–4745. DOI: `10.24963/ijcai.2019/658`. URL: `https://doi.org/10.24963/ijcai.2019/658`.

[11] Shilin He, Jieming Zhu, Pinjia He and Michael R. Lyu. 'Experience Report: System Log Analysis for Anomaly Detection'. In: *2016 IEEE 27th International Symposium on Software Reliability Engineering (ISSRE)*. 2016, pp. 207–218. DOI: `10.1109/ISSRE.2016.21`.

[12] Aurélien Géron. *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow*. 2nd. Sebastopol, CA: O'Reilly Media, 2019. ISBN: 978-1-492-03264-9. URL: `https://www.oreilly.com/library/view/hands-on-machine-learning/9781492032632/`.

Appendix A Structure for the research plan

A research plan can be reported according to the next structure. The order of the items is important.

Introduction

The topic is introduced on general level. The context of the research is described and the research problem is explained and justified. The problem is situated in its larger environment. Note references when needed. The researcher may reason the topic also by describing his or her personal motivation.

Research problem and research methods

The problem under study is explained as explicitly as possible. The research problem can be divided into sub problems or presented as hypotheses. The research methods and analysis are described.

Limitations

The planned limitations and known shortcomings are reported. The reasons for them – if known – are explained from the viewpoint of the current research.

Preliminary earlier research

The prior literature is presented briefly with full sentences. All required references are included. Its relevance in the current research is described and limitations recognised in prior research are identified if possible. List of main prior literature in relation to the background theory Main background references are listed in the required format (APA).

Lähteet

Timetable

A plan to describe the planned research related to calendar time. It is recommended that the plan is discussed with supervisor to ensure enough milestones for checking thoroughly the status of the thesis.

Preliminary structure of contents

1. Introduction
2. Glossary
3. Prior research
    (a) First
    (b) Second
        Subsecond
4. Sources