Journal of Universal Computer Science, vol. 29, no. 11 (2023), 1250-1253
submitted: 11/9/2023, accepted: 30/9/2023, appeared: 28/11/2023 CC BY-ND 4.0

Traceable Use of Emerging Technologies in Smart Systems

J.UCS Special Issue

Wolfram Luther
(University of Duisburg-Essen, Duisburg, Germany,
https://orcid.org/0000-0002-1245-7628, wolfram.luther@uni-due.de)

Gregor Schiele
(University of Duisburg-Essen, Duisburg, Germany,
https://orcid.org/0000-0003-4266-4828, gregor.schiele@uni-due.de)

Keywords: AI-based data mining, application troubleshooting, authentication, automated root cause analysis (RCA), BRCA1/2 mutation probability, cloud infrastructures, cloud-native applications, cyclical patterns, data fusion & integration, Dempster-Shafer evidence theory (DST), distance learning, distributed tracing, explainable AI (XAI), hypothesis testing, interval analysis, key performance indicators (KPI), machine learning (ML), project-based learning, RED metrics, rule-induction systems, sales goals and operation planning, security, social networks, spatio-temporal data, video activity report, work engagement
Categories: B.8, E.2, E.4, G.1, G.3, H.3.1, H.3.2, H.3.3, H.3.7, H.4.2, H.5.1, H.m, I.1.2, I.2, I.2.6, I.5.3 I.5.4, I.6.4, J.3, K.6, L.1.2. L.3.6
DOI: 10.3897/jucs.112577

This volume presents a selection of invited papers from the 3rd Workshop on Collaborative Technologies and Data Science in Smart City Applications (CODASSCA 2022): *From Data to Information and Knowledge*, held in Yerevan, Armenia, August, 23-25, and further articles from a free call for papers JUCS-CODASSCA-2023 published by Easychair. The workshop continues the cooperation between the University of Duisburg-Essen (UDE) and the American University of Armenia (AUA) funded by the German Academic Exchange Service (DAAD) and the German Research Foundation (DFG). The workshop took place together with a one-week summer school on the topic *Enhancements of Deep Learning for Intelligent Applications and the Connected Society*.

In two rounds of review, 15 of the papers submitted in the three formats of extended abstract, short paper, and full paper were selected by the program committee and published in conference proceedings *Data Science, Human-Centered Computing, and Intelligent Technologies* released with a CC BY license by Logos, Berlin 2022, ISBN 978-3-8325-5520-7, https://doi.org/10.30819/5520.

Authors submitted significantly extended and improved versions of their contributions to be considered for a J.UCS special issue *Traceable Use of Emerging Technologies in Smart Systems*. These articles are grouped in two thematic areas:

- *Data science with intelligent technologies;*
- *Human-centered computing with intelligent technologies.*

There was also a J.UCS open call so that any author could submit papers on the highlighted subjects. The invitation to review the 20 contributions received was accepted by 22 experts, and, after three rounds, eight articles were finally accepted for publication in the special issue.

This special volume aims to show ways in which scientists and users can jointly move beyond the current practice of publishing scientific results to more traceable methods and results that facilitate understanding of the goals defined and technologies used to achieve the results and their explainability and validation, that is, to move from "results" to "Xresults" (i.e. the traceable results and explainable technologies used).

The paper written by members of the Armenian National Academy of Sciences is about the two stages of tracing. First, Bob decides that the data is a stegotext and uses the extraction algorithm to identify the hidden message. Then, he uses the key to decide whether it was Eve or Alice who was active.

A group of Armenian and American scientists and developers discuss the application of rule-learning approaches to automated, real-time, and intelligent root cause analysis (RCA) in cloud applications. In the investigation, they use the terms tracing and traces in the sense in which they are listed in, for example, the Oxford Dictionary. A single trace shows an individual request passage through the microservices. Systems trace traffic passing through a malfunctioning microservice to identify and remediate performance degradation issues, finding or describing their origin, development and duration in terms of trace types, spans, and tags.

A group of Armenian, German, and American researchers and R&D specialists deal with ML approaches that help to automate the management of complex systems such as virtual machines, hosts, and datastores by monitoring millions of time series metrics, terabytes of logs, and application traces to capture a high-resolution “image” of the entire stack. Recently, self-diagnostics for issues with such intelligent monitoring and analytics solutions has become another fundamental problem in customer environments. They require time-intensive expert analysis of these traces.

Researchers from Algerian universities use word embedding vectors and multi-grained scanning—a sophisticated extension of text tracing and tracking with sliding windows, Deep Random Forest, and contextual embedding model AraBert—to detect and remove hate speech in Arabic tweets.

Researchers from the Japanese university of Tsukuba consider a prototypical web-based communication system to enable teachers to track student engagement in remote environments. Students upload video activity reports via the system. By automatically fetching a video activity report and analyzing the content of the video, the system can automatically identify students whose engagement is declining.

German researchers from Wismar Applied University describe the requirements for traceable open-source data retrieval to bound mutation probabilities in two tumor-suppressor genes responsible for hereditary breast and/or ovarian cancer. They use a Dempster-Shafer model to consider the family history and first age of onset, as well as epistemic uncertainty.

Chilean scientists and Japanese R&D experts introduce retail sales performance indicators that should be regularly traced and analyzed to identify improvement opportunities and make informed decisions to optimize sales performance. This approach allows users to choose an optimized indicator prediction model for each retail store. Users can use a web tool to estimate improvements in the indicators based on desired sales goals.

Researchers from Chile and the UK deal with periodic pattern mining and look for efficient algorithms to find cyclical patterns in spatio-temporal databases.

Here is a more detailed overview of the contributions:

Data science with intelligent technologies

Two-Stage Optimal Hypotheses Testing for a Model of a Stegosystem with an Active Adversary by Mariam Haroutunian, Parandzem Hakobyan, and Arman Avetisyan.

This paper considers an information-theoretic model of a stegosystem with an active adversary. Unlike a passive adversary, an active adversary could modify the data, whether a covertext or a stegotext, sent by the legitimate transmitter over a public channel. The receiver's task is to decide whether the communication is a covertext or a stegotext and, in case of a stegotext, to further decide whether the message was sent by a legitimate sender or by an adversary. From the receiver's point of view, the authors suggest a two-stage statistical hypothesis testing approach that is logarithmically asymptotically optimal. This approach addresses the functional dependence of errors in both stages.

Distributed Tracing for the Troubleshooting of Native Cloud Applications via Rule-Induction Systems by Arnak Poghosyan, Ashot Harutyunyan, Naira Grigoryan, and Clement Pang.

The authors of this paper consider rule-induction classification methods for the root cause analysis of performance degradations of native cloud application problems. In their experiments, the researchers considered various approaches to determine their suitability for explaining performance degradation and unveiled the main benefits of each method.

Challenges and Experiences in Designing Interpretable KPI-diagnostics for Cloud Applications by Ashot N. Harutyunyan, Arnak V. Poghosyan, Lilit Harutyunyan, Nelli Aghajanyan, Tigran Bunarjyan, and A.J. Han Vinck.

The authors describe the existing technology challenges and their experiences while designing problem root cause analysis mechanisms that are automatic, application agnostic, and interpretable by human operators. They focus on diagnosis of cloud ecosystems through the Key Performance Indicators (KPI). Those indicators are utilized to build automatically labeled data sets and train explainable AI models for identifying conditions and processes "responsible" for misbehaviors. Experiments on a large time series data set from a cloud demonstrate that those approaches are effective in obtaining models that explain unacceptable KPI behaviors and localize sources of issues.

Deep Random Forest and AraBert for Hate Speech Detection from Arabic Tweets by Kheir Eddine Daouadi, Yaakoub Boualleg, and Oussama Guehairia.

This paper addresses the problem of hate speech detection in Arabic tweets. It proposes a method known as Contextual Deep Random Forest (CDRF) by combining Arabic contextual embeddings with Deep Random Forests. The proposed CDRF improves the accuracy rate of hate speech detection and outperforms existing classification approaches.

Human-centered computing with intelligent technologies

Using Video Activity Reports to Support Remote Project-based Learning by Kosuke Sasaki, Zhen He, and Tomoo Inoue.

This paper reports on the authors' experience with video activity reports in a project-based remote-learning activity in higher education. Characteristics of the video reports (length, pauses, negative statements) are contrasted with self-reported engagement scores using a minimalist instrument (UWES-3 questionnaire) developed in a work-related context. The findings are conceived as steps towards designing an AI-enabled support system.

Towards a Traceable Data Model Accommodating Bounded Uncertainty for DST Based Computation of BRCA1/2 Mutation Probability with Age by Lorenz Gillner and Ekaterina Auer.

The authors present a novel uncertainty-aware model for determining the risk of cancer diseases. Their model is based on interval extensions of Dempster-Shafer evidence theory and applied successfully to data about breast cancer available in open-access publications and databases.

Retail Indicators Forecasting and Planning by Nelson Baloian, Jonathan Frez, José A. Pino, Cristóbal Fuenzalida, Sergio Peñafiel, Belisario Panay, Gustavo Zurita, and Horacio Sanson.

The article presents a methodology for forecasting plausible business goals for a particular retail store for the next two months. This methodology supports managers in planning sales goals and operation.

Efficiently Finding Cyclical Patterns on Twitter Considering the Inherent Spatio-Temporal Attributes of Data by Claudio Gutiérrez-Soto, Patricio Galdames, and Daniel Navea

This paper introduces HashCycle, a new and efficient algorithm for finding cyclical patterns over spatio-temporal databases.

The editors would like to thank the authors and express their gratitude to the AUA and Dr. Aram Hajian, dean of the College of Engineering, for hosting the CODASSCA workshop series, to the J.UCS consortium for accepting the special issues in the Journal of Universal Computer. Science, and to Johanna Zeisberg and the publishing team for their kind assistance throughout the process of writing and compiling the issue.

Thanks also to the German Research Foundation (DFG), the German Academic Exchange Service (DAAD), the UDE publication fund, and the sponsors involved for funding our common activities since 2015. Without the help of the CODASSCA 2022 editors and their ongoing encouragement, the present volume would not have been possible.

The guest editors are especially grateful to the anonymous reviewers who spent a significant amount of time reading the articles and making substantive suggestions for improvement., We are particularly indebted to Dr. Yanling Chen, Prof. Nelson Baloian, and Dr. Ashot Harutyunyan for their invaluable help during the preparation of this special issue.

Wolfram Luther
Gregor Schiele
Duisburg
September 2023

Journal of Universal Computer Science, vol. 29, no. 11 (2023), 1254-1273
submitted: 19/5/2023, accepted: 30/9/2023, appeared: 28/11/2023 CC BY-ND 4.0

Two-Stage Optimal Hypotheses Testing for a Model of Stegosystem with an Active Adversary

Mariam Haroutunian
(Institute for Informatics and Automation Problems of NAS of RA, Yerevan, Armenia
https://orcid.org/0000-0002-9262-4173, armar@sci.am)

Parandzem Hakobyan
(Institute for Informatics and Automation Problems of NAS of RA, Yerevan, Armenia
https://orcid.org/0000-0002-5056-9591, par_h@iiap.sci.am)

Arman Avetisyan
(Institute for Informatics and Automation Problems of NAS of RA, Yerevan, Armenia
https://orcid.org/0000-0002-0434-2767, armanavetisyan1997@gmail.com)

Abstract: We study the information-theoretic model of stegosystem with an active adversary, where unlike a passive adversary he can not only read but also write. The legitimate sender as well as the adversary can embed or not a message in the sending data. The receiver's first task is to decide whether the communication is a covertext, data with no hidden message, or a stegotext, modified data with a hidden secret message. In case of stegotext, the receiver's second task is to decide whether the message was sent by a legitimate sender or from an adversary. For this purpose an authenticated encryption from the legitimate sender is considered.

In this paper we suggest two-stage statistical hypothesis testing approach from the receivers point of view. We propose the logarithmically asymptotically optimal testing for this model. As a result the functional dependence of reliabilities of the first and second kind of errors in both stages is constructed. A comparison of overall error probabilities with the situation of one stage hypotheses testing is discussed and the behaviour of functional dependences of reliabilities are illustrated.

Keywords: Security, Hypotheses testing, Authentication.
Categories: E.4
DOI: 10.3897/jucs.112913

1 Introduction

The aim of steganography is communicating messages by hiding them within other data thereby creating a covert channel. By standard terminology of information hiding [Pfitzmann 1996] the legitimate users are Alice and Bob, who wishes to communicate over a public channel, such that the presence of hidden message must be unnoticed to an adversary (Eve).

Various models with various tasks have been studied [Katzenbeisser et al. 2002], [Hopper et al. 2002], [Von Ahn and Hopper 2004], [Backes and Cachin 2005], [Dedic et al. 2009], [Liśkiewicz et al. 2011], [Augot et al. 2011]. We are interested in information-theoretic investigations studied in many papers including [O'Sullivan et al. 1998], [Moulin and O'Sullivan 2003], [Cachin 2004], [Mittelholzer 2000], [Wang and Moulin 2008], [Shikata and Matsumoto 2008], [Balado and Haughton 2018].

In [Cachin et al. 1998], [Cachin 2004] Cachin first proposed an information-theoretic model of steganography with passive adversary (who has read-only access to the public channel) (Fig. 1). Alice can be inactive and send a covertext C without hidden information or be active and send stegotext S. Bob has the exctracting algorithm, but Eve does not know if Alice was active or not. Hence, Eve must solve the problem of Hypothesis testing.

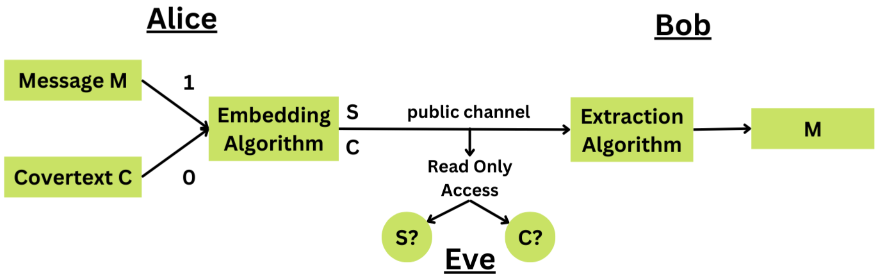


Figure 1: The model of stegosystem with passive attacks.

An extended information-theoretic model for steganography with active attacks (where adversary can read and write a message over an insecure channel) was proposed and studied in [Shikata and Matsumoto 2008]. More specifically, the authors showed a generic construction of secure stegosystems by using almost unbiased functions and secure authenticated encryption with random ciphertexts in the model with active adversaries in unconditional setting. The problem of information-theoretically secure authenticated encryption is addressed in [Alomair and Poovendran 2009].

In this paper we consider an information-theoretic model of a stegosystem with active attacks (Fig. 2), we propose and study the problem of optimal hypothesis testing, which will be described later in this section.

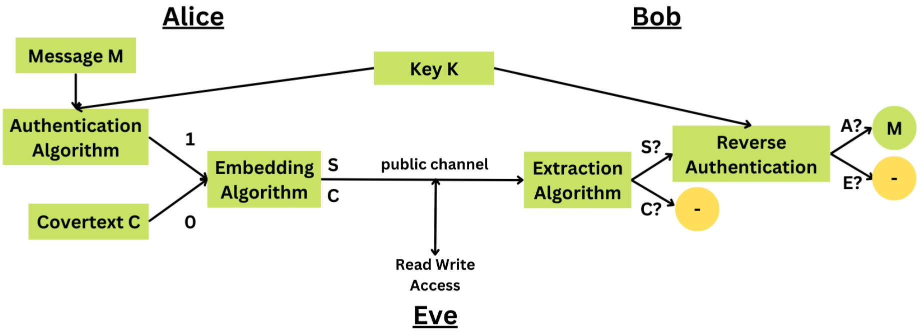


Figure 2: The model of stegosystem with active adversary.

Adversary has an access to a read and write public channel and is able to analyze and modify data. Alice as well as Eve can be either active or passive, i.e. can embed or not a message in the sending data. Bob's first task is to decide whether the received data X is a covertext C, data with no hidden message, or stegotext S, modified data with a hidden secret message M. In case of deciding that the obtained data is stegotext, Bob has the extraction function, and the second task for Bob is to decide whether the extracted message was sent by Alice or Eve. For this purpose an authenticated encryption of message M with secret key K is considered. Depending on applications this encryption except authentication can include also secrecy requirements of hidden message.

Covertext is generated by a source according to a distribution P_C, stegotext has a distribution P_S according to a certain embedding function. The distribution of secret key we denote by P_K. We assume that Eve knows all these distributions. For the authenticated encryption Alice generates the encrypted message according to P_{MK} and Eve can generate a message with distribution $P_M P_K$.

We suggest two-stage statistical hypotheses testing approach from receivers point of view. On the first stage Bob has to decide if the data was generated according to P_C or P_S. In the case when Bob decides that stegotext is obtained, after extracting the secret message, on the second stage Bob has to decide if the message was generated according to P_{MK} or $P_M P_K$. Further, we substantiate the advantages of our approach.

The paper is organized as follows. In the next section the considered problem and the related art are presented. The main notations and definitions are given in the section 3 and the results are formulated in section 4. Some discussions on partial cases and reccomendations are provided in section 5. In section 6 some illustrations of the functional behaviour is presented. The conclusion remarks are in section 7. The detailed proofs of the theorems are placed in the appendix.

2 Problem Statement

In classical statistical hypothesis testing problem a statistician makes decision on which of the two proposed hypotheses H_1 and H_2 must be accepted based on data samples. This decision is made on the certain procedure which is called test. Due to randomness of the data the result of this decision may lead to two types of errors: the fist type is called the error for accepting H_2 when H_1 is true and the second type error for accepting H_1 when H_2 is true. In such problems the aim is to find such a test, that reduces both types of errors as much as possible. The complexity of the task is that the two types of errors are interconnected, when the one is reduced the other one can get increased.

Another problem related with information theory is the case of a tests sequence, where the error probabilities are decreasing exponentially as 2^{-NE}, when the number of observations N is increasing. The exponent of error probability E is called *reliability*. In the case with two hypotheses both reliabilities corresponding to two possible error probabilities could not increase simultaneously. It is an accepted way to fix the value of one of the reliabilities and try to make the tests sequence get the greatest value of the remaining reliability. Such a test is called *logarithmically asymptotically optimal* (LAO). The publications [Hoeffding 1965], [Csiszár and Longo 1971] , [Blahut 1974], [Tusnady 1977], [Longo and Sgarro 1980], [Birgé 1981],[Haroutunian 1989], [Haroutunian 1990] and [Haroutunian et al. 2008] are devoted to this problem, particularly, the problem of multiple hypotheses LAO testing was investigated in [Haroutunian 1989], [Haroutunian 1990], [Haroutunian et al. 2008]. Multiple hypotheses testing was proposed in many studies as an effective framework for analyzing problems in various domains, such as performance evaluation of biometric systems (see [Willems et al 2003], [Harutyunyan et al. 2011], [Yagi and Hirasawa 2022]. This framework enables looking into the underlying problems from the information-theoretic perspectives and optimal achievability bounds of error probability trade-offs while treating observations data emitted from classical models of information sources like Discrete Memoryless Source or Arbitrarily Varying Source (AVS) [Harutyunyan and Han Vinck 2006], [Grigoryan and Harutyunyan 2015], [Grigoryan et al. 2011].

The problem of LAO testing of statistical hypotheses for the steganography model with a passive adversary (Fig. 1) was solved in [Haroutunian et al. 2018]. In that model the adversary's task was to distinguish the covertext from stegotext. The functional dependence of the reliabilities of the first and the second kind errors was given.

In this paper we suggest two stage logarithmically asymptotically optimal testing of the legal receiver for the steganographic model with active adversary. At the first stage Bob decides whether a covertext or a stegotext is received. If at the first stage Bob decides that the data is a stegotext, then he uses the extraction algorithm to get the hidden message and using the key at the second stage he decides whether Eve or Alice was active.

We study the functional dependence of reliabilities of the first and second kind of errors of optimal tests in both stages. The proof of the result for first stage is similar to the result suggested in [Haroutunian et al. 2018], where the problem of LAO testing of statistical hypotheses for the steganography model with a passive adversary is solved by the method of types [Csiszár 1998]. For the second stage the approach studied in [Maurer 2000] was useful.

The results of this paper partially were reported at the CODASSCA Workshop [Haroutunian et al. 2022]. Here we introduce the full version, i.e with added proofs, discussions on advantages of our approach and illustrations of the theoretical dependences.

3 Notations and Definitions

Here we present some necessary characteristics and results of information theory [Blahut 1987], [Cover and Thomas 2006]. We denote finite sets by script capitals. The cardinality of a set $\mathcal{X}$ is denoted as $|\mathcal{X}|$. We denote random variables (RV) by X, S, C, K, M. Probability distributions (PD) are denoted by P, P_C, P_S, P_M, P_K, Q and P_{MK}.

Let PD of RV K and M be

$$P_K \triangleq \{P_K(k), \quad k \in \mathcal{K}\},$$

$$P_M \triangleq \{P_M(m), \quad m \in \mathcal{M}\},$$

and the joint PD of RVs M and K be

$$P_{MK} \triangleq \{P_{MK}(m,k), \ m \in \mathcal{M}, \ k \in \mathcal{K}\}.$$

The Shannon entropy $H_P(X)$ of RV X with PD $P \triangleq \{P = P(x), \ x \in \mathcal{X}\}$ is:

$$H_P(X) \triangleq -\sum_{x\in\mathcal{X}} P(x)\log P(x).$$

The mutual information of RV M and K equals:

$$I_{P_{MK}}(M;K) \triangleq \sum_{m\in\mathcal{M},\ k\in\mathcal{K}} P_{MK}(m,k)\log\frac{P_{MK}(m,k)}{P_M(m)P_K(k)}.$$

It is important to note that

$$P_M(m) = \sum_{k\in\mathcal{K}} P_{MK}(m,k),$$

$$P_K(k) = \sum_{m\in\mathcal{M}} P_{MK}(m,k).$$

The joint entropy of RVs M and K is the following:

$$H_{P_{MK}}(M,K) \triangleq -\sum_{m\in\mathcal{M},\ k\in\mathcal{K}} P_{MK}(m,k)\log P_{MK}(m,k).$$

We use the notion of divergence (Kullback-Leibler information or "distance") defined on two PDs, say P_C and P_S, on $\mathcal{X}$ as:

$$D(P_C||P_S) \triangleq \sum_{x\in\mathcal{X}} P_C(x)\log\frac{P_C(x)}{P_S(x)}.$$

The divergence of joint PDs $Q \triangleq \{Q = Q(m,k),\ \ m\in\mathcal{M},\ k\in\mathcal{K}\}$ and P_{MK} on $(\mathcal{M}\times\mathcal{K})$ is:

$$D(Q||P_{MK}) \triangleq \sum_{m\in\mathcal{M},k\in\mathcal{K}} Q(m,k)\log\frac{Q(m,k)}{P_{MK}(m,k)}.$$

The space of all joint PDs on finite set $\mathcal{M}\times\mathcal{K}$ we denote by

$$\mathcal{Q}(\mathcal{M}\times\mathcal{K}) \triangleq \{Q : Q = Q(m,k), m\in\mathcal{M},\ \ k\in\mathcal{K}\}.$$

When RV M and K are independent, then

$$\begin{aligned} D(Q||P_{MK}) &= D(Q||P_M P_K) \\ &= \sum_{m\in\mathcal{M},k\in\mathcal{K}} Q(m,k)\log\frac{Q(m,k)}{P_M(m)P_K(k)}. \end{aligned}$$

In particular, the divergence of PDs P_{MK} and $P_M P_K$ is the mutual information:

$$\begin{aligned} D(P_{MK}||P_M P_K) &\triangleq \sum_{m\in\mathcal{M},k\in\mathcal{K}} P_{MK}(m,k)\log\frac{P_{MK}(m,k)}{P_M(m)P_K(k)} \\ &= I_{P_{MK}}(M;K). \end{aligned}$$

For our investigations we use the method of types, [Haroutunian et al. 2008], [Csiszár 1998], [Csiszár and Körner 1981], the essence of which is to partition the set of all same length vectors into classes according to their empirical distributions.

The type $P_{\mathbf{x}}$ of a vector $\mathbf{x} = (x_1, ..., x_L) \in \mathcal{X}^L$ is a PD (the empirical distribution)

$$P_{\mathbf{x}} = \left\{ P_{\mathbf{x}}(x) = \frac{N(x|\mathbf{x})}{L}, x \in \mathcal{X} \right\},$$

where $N(x|\mathbf{x})$ is the number of repetitions of symbol x in vector $\mathbf{x}$. We denote by $\mathcal{P}^L(\mathcal{X})$ the set of all types of vectors in $\mathcal{X}^L$ for given L and the set of vectors $\mathbf{x}$ of type $P_{\mathbf{x}}$ is denoted by $\mathcal{T}^L_{P_{\mathbf{x}}}(X)$.

The joint type of vectors $\mathbf{m} = (x_1, ..., x_N) \in \mathcal{M}^N$ and $\mathbf{k} = (k_1, ..., k_N) \in \mathcal{K}^N$ $Q_{\mathbf{m},\mathbf{k}}$ a PD (the empirical distribution)

$$Q_{\mathbf{m},\mathbf{k}} = \left\{ Q_{\mathbf{m},\mathbf{k}} = \frac{N(m,k|\mathbf{m},\mathbf{k})}{N}, m \in \mathcal{M},\ \ k \in \mathcal{K} \right\},$$

where N(m,k|**m, k**) is the number of repetitions of symbols pair (m, k) in the pair of vectors $(\mathbf{m}, \mathbf{k})$. The set of all joint types of vector pairs $(\mathbf{m}, \mathbf{k})$ in $(\mathcal{M} \times \mathcal{K})^N$ for given N is denoted by $\mathcal{Q}^N(\mathcal{M} \times \mathcal{K})$ and the set of vector pairs $(\mathbf{m}, \mathbf{k})$ of type $Q_{\mathbf{m},\mathbf{k}}$ is denoted by $\mathcal{T}^N_{Q_{\mathbf{m},\mathbf{k}}}(M, K)$.

The following well known inequalities [Csiszár and Körner 1981] we use in the proofs of our results :

$$| \mathcal{P}^L(\mathcal{X}) | \leq (L+1)^{|\mathcal{X}|}, \tag{1}$$

$$| \mathcal{Q}^N(\mathcal{M} \times \mathcal{K}) | \leq (N+1)^{|\mathcal{M}||\mathcal{K}|}, \tag{2}$$

for any type $P \in \mathcal{P}^L(\mathcal{X})$

$$(L+1)^{-|\mathcal{X}|} \exp\{LH_P(X)\} \leq | \mathcal{T}^L_P(X) | \leq \exp\{LH_P(X)\}, \tag{3}$$

for any type $Q \in \mathcal{Q}^N(\mathcal{M} \times \mathcal{K})$

$$(N+1)^{-|\mathcal{M}||\mathcal{K}|} \exp\{NH_Q(M,K)\} \leq | \mathcal{T}^N_Q(M,K) | \leq \exp\{NH_Q(M,K)\}. \tag{4}$$

The method of types is one of the important technical tools in Information Theory.

4 Formulation of Results

First stage. At the first stage, from the received data $\mathbf{x} = (x_1, ..., x_L)$, $\mathbf{x} \in \mathcal{X}^L$, Bob must decide whether it is a covertext or a stegotext. Hence, Bob must accept one of the two hypotheses

$$H_1 : P = P_S \quad \{\text{data is a stegotext}\}$$

$$H_2 : P = P_C \quad \{\text{data is a covertext}\}.$$

The procedure of decision making is a non-randomized test φ_L, which can be defined by partition of the set of possible messages $\mathcal{X}^L$ on two disjoint subsets $\mathcal{A}^L_i$, $i = \overline{1,2}$. The set $\mathcal{A}^L_i$, $i = \overline{1,2}$ contains all data $\mathbf{x}$ for which the hypothesis H_i is adopted.

The first kind error probability, which is the probability of the rejection of the correct hypothesis H_1 is the following:

$$\alpha_{2|1}(\varphi_L) = P_S^L(\mathcal{A}_2^L).$$

The second kind error probability, which is the probability of the erroneous acceptance of hypothesis H_1 is defined as follows:

$$\alpha_{1|2}(\varphi_L) = P_C^L(\mathcal{A}_1^L).$$

The error probability exponents, called “reliabilities” of the infinite sequence of tests φ, are defined respectively as follows:

$$E_{2|1}^I(\varphi) \overset{\triangle}{=} \underline{\lim}_{L\to\infty} -\frac{1}{L}\log\alpha_{2|1}(\varphi_L),$$

$$E_{1|2}^I(\varphi) \overset{\triangle}{=} \underline{\lim}_{L\to\infty} -\frac{1}{L}\log\alpha_{1|2}(\varphi_L).$$

As defined in [Birgé 1981] the sequence of tests φ^* is called **logarithmically asymptotically optimal** (LAO) if for given positive value of $E_{2|1}^I$ the maximum possible value is provided for $E_{1|2}^I$.

The procedure for creating an optimal decision rule is similar to [Haroutunian et al. 2018]. Our first result, that is the functional dependence of the reliabilities of the first and second kind of errors is given by the following theorem.

Theorem 1. *For given*

$$0 < E_{2|1}^I < D(P_C||P_S) \tag{5}$$

there exists a LAO sequence of tests, the reliability $E_{1|2}^{*,I}$ *of which is defined as follows:*

$$E_{1|2}^{*,I} = E_{1|2}^{*,I}(E_{2|1}^I) = \inf_{P:\ D(P||P_S)\le E_{2|1}^I} D(P||P_C). \tag{6}$$

When $E_{2|1}^I \geq D(P_C||P_S)$, then $E_{1|2}^{*,I}$ is equal to 0 .

Thus, for a given reliability of incorrectly rejecting the stegotext, we get the maximal reliability of wrongly accepting the stegotext.

Comment 1: Unlike model considered in [Cachin 2004], [Haroutunian et al. 2018], here Bob has no additional information about whether Alice is active or passive. Therefore, considered stegosystem should not be *perfectly secure*, because otherwise Bob cannot find out that he has received a covertext or a stegotext. Hence, we assume that for distributions P_C and P_S, $D(P_C||P_S) > 0$.

Comment 2. For given $E_{2|1}^I \in (0, D(P_C||P_S))$ the following holds:

$$E_{1|2}^{*,I} < D(P_S||P_C).$$

If at the first stage Bob accepts the hypothesis H_1, which means that he decides that the data is a stegotext, then he uses the extraction algorithm to get the hidden message

$\mathbf{m} = (m_1, m_2, ..., m_N)$.

Second stage. After the extraction using key sequence $\mathbf{k} = (k_1, k_2, ..., k_N)$ Bob has to decide whether Eve or Alice sent him that message. So he moves on to the second stage of hypothesis testing:

$$H_1: \quad Q = P_{MK}(m,k) \qquad \{\text{there was no attack}\}$$

$$H_2: \quad Q = P_M(m)P_K(k) \qquad \{\text{there was attack}\}.$$

In this case the test Φ_N is defined by partition of the set $(\mathcal{M} \times \mathcal{K})^N$ on two disjoint subsets $\mathcal{B}_l^N$, $l = \overline{1,2}$. The set $\mathcal{B}_1^N$ contains all data pairs $(\mathbf{m}, \mathbf{k})$ for which the hypothesis H_1 is adopted, which in our context means that message $\mathbf{m}$ is sent from Alice. Correspondingly, the set $\mathcal{B}_2^N$ contains all pairs $(\mathbf{m}, \mathbf{k})$ for which the hypothesis H_2 is adopted, i.e. Bob decides that message is sent from Eve.

The probabilities of errors of the first and second kind by analogy to the case of the first stage are defined as follows:

$$\alpha_{2|1}^{II}(\Phi_N) = P_{MK}^N(\mathcal{B}_2^N), \text{ (the first kind error probability)}$$

$$\alpha_{1|2}^{II}(\Phi_N) = (P_M P_K)^N(\mathcal{B}_1^N), \text{ (the second kind error probability).}$$

The error probability exponents of the infinite sequence of tests Φ, are defined respectively as follows:

$$E_{i|j}^{II}(\Phi) \stackrel{\triangle}{=} \varliminf_{N\to\infty} -\frac{1}{N}\log \alpha_{i|j}^{II}(\Phi_N), \ i \neq j, \ \ i,j = \overline{1,2}.$$

The second kind error probability in Bob's decision essentially coincides with the probability of Eve's succeeding. Hence, the maximum value of $E_{1|2}^{II}$ guarantees that the attacker will fail.

As in the First Stage, for given positive value $E_{2|1}^{II}$ we constructed the LAO sequence of tests Φ^* and the dependence of maximal value $E_{1|2}^{II}$ from $E_{2|1}^{II}$ is provided in the following theorem.

Theorem 2. *For given*

$$0 < E_{2|1}^{II} < D(P_M P_K || P_{MK}) \tag{7}$$

there exists a LAO sequence of tests, the reliability $E_{1|2}^{*,II}$ of which is defined as follows:

$$E_{1|2}^{*,II}\left(E_{2|1}^{II}\right) = \inf_{Q:\, D(Q||P_{MK}) \leq E_{2|1}^{II}} D(Q||P_M P_K). \tag{8}$$

When $E_{2|1}^{II} \geq D(P_M P_K||P_{MK})$, then $E_{1|2}^{*,II}$ is equal to 0 .

Comment 3. For given $E_{2|1}^{II} \in (0, D(P_M P_K||P_{MK}))$ the following holds:

$$E_{1|2}^{*,II} < D(P_{MK}||P_M P_K) = I_{P_{MK}}(M;K).$$

The proofs of the theorems are given in the appendix, the brief outline of which is the following. Decision regions of the first and second stages, which are the disjoint subsets of the corresponding sample spaces, are constructed by comparing the divergance of the sample type and the distribution of the first hypothesis with the given reliability of the first error. When divergance is less than or equal to the given reliability, the first hypothesis is accepted, otherwise, the second hypothesis is accepted. According to the properties of types, the optimality of such sample space divisions is substantiated and the dependence of reliabilities is established.

5 Discussions

Obviously, by skipping the first stage of hypotheses testing, that is, using the extraction algorithm for all the data received, Bob can gain in error probability of the first stage at loss in time. Two questions arise.

- How much does the total error probability of the two-stage model differ from the one-stage case?
- Is it possible to propose situations, where the total error probability of the two-stage model is equal to the one-stage case, therefore, the gain will be in time without loss in the error probabilities?

Let us consider the total error probabilities and exponents in two-stage approach. First notice that $\alpha^{I}_{1|2}$ is not a principal error, because deciding that the data is a stegotext, while it is a covertext, will be discovered on extracting phase and hence, this error can be ignored.

If we denote by $\alpha_{2|1}$ the probability of Alice's failure, i.e. that Bob erroneously rejects the useful information sent by Alice, then

$$\alpha_{2|1} = \alpha^{II}_{2|1} + \alpha^{I}_{C|S\&Alice} \leq \alpha^{II}_{2|1} + \alpha^{I}_{2|1} \leq 2\max(\alpha^{II}_{2|1}, \alpha^{I}_{2|1}),$$

where $\alpha^{I}_{C|S\&Alice}$ is the error probability when data includes message sent by Alice, but Bob rejects it in the first stage deciding it as covertext.

For the corresponding reliability $E_{2|1}$ in one-stage scenario the following inequality takes place

$$E_{2|1} \geq \min(E^{II}_{2|1}, E^{I}_{2|1}).$$

On the other hand

$$\alpha_{2|1} = \alpha^{II}_{2|1} + \alpha^{I}_{C|S\&Alice} \geq \alpha^{II}_{2|1}.$$

There is also one principal error probability $\alpha_{1|2}$, when accepting fake message sent by Eve.

$$\alpha_{1|2} = \alpha^{II}_{1|2} + \alpha^{I}_{1|2} = \alpha^{II}_{1|2},$$

because as it was mentioned above we can ignore $\alpha^{I}_{1|2}$ as a not principal error.

Therefore, the overall reliability $E_{1|2}$ is equal to $E^{II}_{1|2}$.

Thus, for $E_{2|1}$ and $E_{1|2}$ we have the following:

$$\min(E^{II}_{2|1}, E^{I}_{2|1}) \leq E_{2|1} \leq E^{II}_{2|1}, \tag{9}$$

$$E_{1|2} = E^{II}_{1|2}. \tag{10}$$

Now through a pair $(E^{I}_{2|1}, E^{II}_{2|1})$ of given reliabilities of the first and second stages, we can make a judgement about the total reliabilities $(E_{2|1}, E_{1|2})$.

The pair of total reliabilities $(E_{2|1}, E_{1|2})$ can be also obtained by one-stage testing. In this approach, Bob uses the extraction algorithm to get the hidden message $\mathbf{m} = (m_1, m_2, ..., m_N)$. Obviously the data which are covertexts, are not taken into account during this process. This means that the error probability of the $\alpha^{I}_{C|S\&Alice}$ discussed in the previous scenario is equal to zero. After that he performs hypotheses testing in similar way at the second stage:

$$H_1: \quad Q = P_{MK}(m,k) \qquad \{\text{there was no attack}\}$$

$$H_2: \quad Q = P_M(m)P_K(k) \qquad \{\text{there was attack}\}$$

For optimal testing we have

$$0 < E_{2|1} < D(P_M P_K || P_{MK}) \tag{11}$$

and find optimal $E^{*}_{1|2}$:

$$E^{*}_{1|2}\left(E_{2|1}\right) = \inf_{Q:\ D(Q||P_{MK}) \le E_{2|1}} D(Q||P_M P_K). \tag{12}$$

We are interested in the situations where for the indicated reliabilities we will get the same result with a two-stage approach as with the one-stage case. We shall show that such cases exist.

According to (9), for $E^{I}_{2|1} \ge E^{II}_{2|1}$ the reliability $E_{2|1}$ is equal to $E^{II}_{2|1}$. Thus, according to (9),(10), (8) and (12), the results of two and one stage approaches are the same. More specifically:

1. When

   $$D(P_M P_K || P_{MK}) \ge D(P_C || P_S),$$

   then for each

   $$0 < E^{I}_{2|1} < D(P_C || P_S)$$

   Bob can choose the reliability $E^{II}_{2|1}$, such that $E^{II}_{2|1} \le E^{I}_{2|1}$, satisfying this condition:

   $$0 < E^{II}_{2|1} \le D(P_C || P_S).$$

   In this case the results of two-stage and one-stage approaches are the same for $E_{2|1} \in \Big(0, D(P_C||P_S)\Big)$.

   Under the condition $D(P_M P_K || P_{MK}) \ge D(P_C || P_S)$ for each

   $$D(P_C || P_S) < E^{II}_{2|1} < D(P_M P_K || P_{MK})$$

   the reliability $E^{II}_{2|1}$ is greater than $E^{I}_{2|1}$, hence the total reliability of the two-stage test $E_{2|1}$ is less than the corresponding one-stage reliability (see (9)).

   In this case the results of two-stage and one-stage approaches are not the same.

2. When

$$D(P_M P_K||P_{MK}) \leq D(P_C||P_S),$$

then for each

$$0 < E^I_{2|1} < D(P_C||P_S)$$

Bob can choose the reliability $E^{II}_{2|1}$, such that

$$0 < E^{II}_{2|1} < D(P_M P_K||P_{MK}), \quad E^{II}_{2|1} \leq E^I_{2|1}.$$

For this case the results of two and one stage approaches are the same always, because the reliability of $E_{2|1}$ for two-stage and one-stage testing stays the same (it is true for all situations discussed above). Moreover, unlike the case 1, the changing ranges of the corresponding reliabilities $E_{2|1}$ for one-stage and two-stage testing are the same.

For all situations discussed above, the reliabilities $E^I_{1|2}$ of two-stage and one-stage testing are equal. This claim is justified by considering (9), (10), (8), (11) and (12).

6 Illustration of Results

To see the behaviour of the functions obtained in Theorem 1 and Theorem 2, here we consider simple examples with illustrations.

Consider the binary set $\mathcal{X}$ and the following distorions are given on $\mathcal{X}$:

$$P_C = \{0.2, 0.8\}, \; P_S = \{0.35, 0.65\}.$$

The values of the following divergences are:

$$D(P_C||P_S) \approx 0.005419, \; D(P_S||P_C) \approx 0.00609.$$

On Fig. 3 the function $E^{*,I}_{1|2}(E^I_{2|1})$ (6) is presented. Here (see (5)),

$$0 < E^I_{2|1} < 0.005419.$$

For the values $E^I_{2|1} \geq 0.005419$ the reliability $E^{*,I}_{1|2}$ equals 0.

Now let the distribution P_{MK} on the set $\mathcal{M} \times \mathcal{K}$ is given by the following matrix

$$P_{MK} = \begin{pmatrix} 0.1, & 0.2 \\ 0.3 & 0.4 \end{pmatrix}.$$

Then from the joint distribution P_{MK} we find

$$P_M = (0.3, 0.7), \; P_K = (0.4, 0.6).$$

The values of the divergences are:

$$D(P_M P_K||P_{MK}) \approx 0.004088, \; D(P_{MK}||P_M P_K) \approx 0.004022.$$

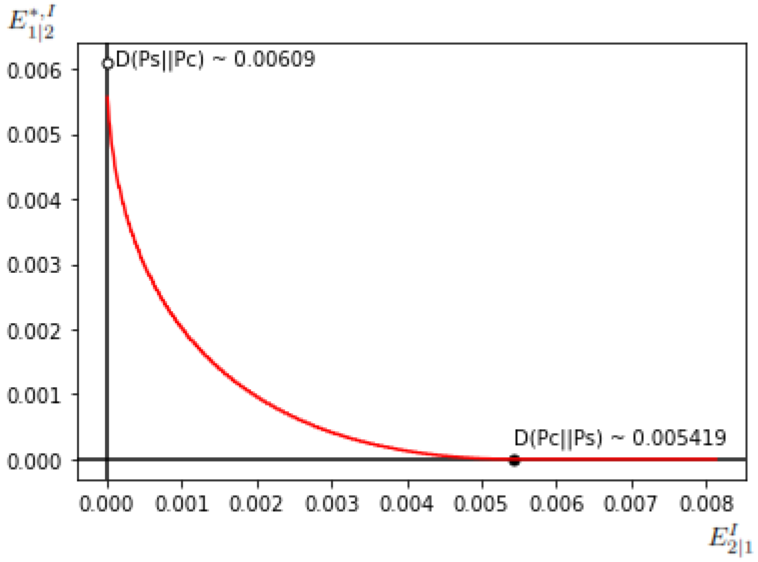


Figure 3: The dependence of reliabilities for the first stage of LAO test.

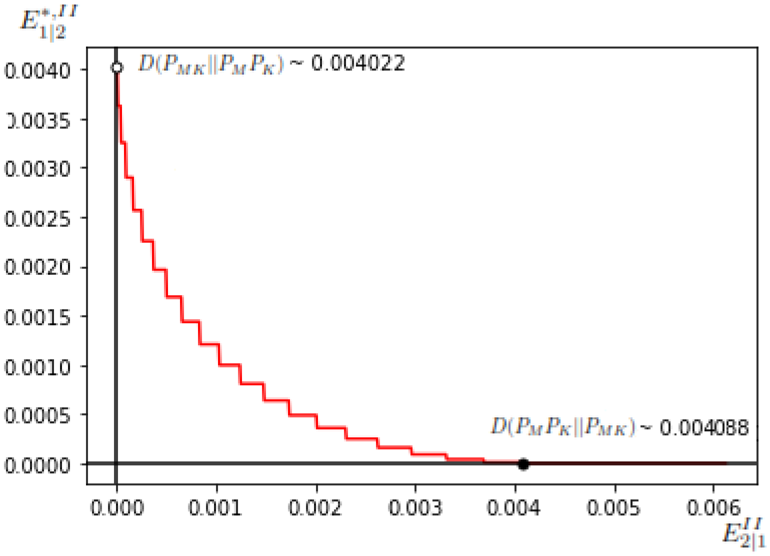


Figure 4: The dependence of reliabilities for the second stage of LAO test.

On Fig. 4 the dependence of reliabilities $E^{*,II}_{1|2}(E^{II}_{2|1})$ of the second stage obtained in

Theorem 2 (see (8), (7)) is presented. As we see, when

$$0 < E^{II}_{2|1} < 0.004088$$

then

$$0 < E^{*,II}_{1|2}(E^{II}_{2|1}) < 0.004022.$$

In the case of one-stage approach the illustration of dependence of the reliabilities $E_{2|1}$ and $E^{*}_{1|2}$ is similar to Fig. 4.

From two-stage testing we can get the same result for total reliabilities $E_{2|1}$ and $E_{1|2}$, if for every $E^{I}_{2|1} \in \big(0,\ 0.005419\big)$ Bob will choose $E^{II}_{2|1} \in \big(0,\ 0.004088\big)$, such that $E^{II}_{2|1} \leq E^{I}_{2|1}$, and vice versa, for every $E^{II}_{2|1} \in \big(0,\ 0.004088\big)$ Bob can find $E^{I}_{2|1} \in \big(0,\ 0.005419\big)$, such that $E^{II}_{2|1} \leq E^{I}_{2|1}$. According to case 2, the scatter plot of all pairs $(E_{2|1}, E_{1|2})$ will have an image like Fig. 4.

7 Conclusions

Two-stage statistical hypothesis testing approach from the receivers point of view in the stegosystem with active adversary is considered. The logarithmically asymptotically optimal testing for this model is analyzed. As a result the functional dependence of reliabilities of the first and second kind of errors in both stages is constructed. The advantages of the two-stage approach are discussed. The elaboration on behavior of derived exponents justifies the theoretical viability of the proposed framework.

In our future work, we will consider using this research to create an e-voting model with added security.

APPENDIX

The proof of Theorem 1.

The proof of the theorem is carried out in the following steps. First we show that for a given number $E^{I}_{2|1}$ we can construct a test. And at the second part, we prove that the constructed test is LAO. Let us consider the following subsets of $\mathcal{X}^L$:

$$\mathcal{B}^L_1 = \bigcup_{P_{\mathbf{x}}:\ D(P_{\mathbf{x}}||P_S) \leq E^{I}_{2|1}} \mathcal{T}^L_{P_{\mathbf{x}}}(X),$$

$$\mathcal{B}^L_2 = \bigcup_{P_{\mathbf{x}}:\ D(P_{\mathbf{x}}||P_S) > E^{I}_{2|1}} \mathcal{T}^L_{P_{\mathbf{x}}}(X).$$

We want to prove, that this division for given $E^{I}_{2|1}$ determines a test φ^*_L. Thus, we are going to validate, that

1. $\mathcal{X}^L = \mathcal{B}^L_1 \cup \mathcal{B}^L_2$;
2. $\mathcal{B}^L_1 \cap \mathcal{B}^L_2 = \emptyset$;

3. $\alpha_{2|1}(\varphi_L^*) \approx 2^{-LE_{2|1}^I}$

The proof of the fist and the second claims are obvious.

Let us prove the third claim, i.e. given $E_{2|1}^I$ is the reliability of error probability $\alpha_{2|1}(\varphi_L^*)$ of tests φ_L^*.

For the proofs we use the known properties of types [Csiszár 1998]:
if $\mathbf{x} \in \mathcal{T}_{P_\mathbf{x}}^L(X)$, then

$$P^L(\mathbf{x}) = \exp\{-L(H_{P_\mathbf{x}}(X) + D(P_\mathbf{x}||P))\}. \tag{13}$$

$$(L+1)^{-|\mathcal{X}|}\exp\{-LD(P_\mathbf{x}||P)\} \leq P^L(\mathcal{T}_{P_\mathbf{x}}^L(X)) \leq \exp\{-LD(P_\mathbf{x}||P)\}. \tag{14}$$

According to (1), (3) and (14), we can estimate $\alpha_{2|1}^I(\varphi_L^*)$ by the following way:

$$\begin{aligned}\alpha_{2|1}(\varphi_L^*) &= P_S^L\left(\mathcal{B}_2^L\right)\\ &= P_S^L\left(\bigcup_{P_\mathbf{x}:D(P_\mathbf{x}||P_S)>E_{2|1}^I} \mathcal{T}_{P_\mathbf{x}}^L(X)\right)\\ &\leq (L+1)^{|\mathcal{X}|}\sup_{P_\mathbf{x}:D(P_\mathbf{x}||P_S)>E_{2|1}^I} P_S^L\left(\mathcal{T}_{P_\mathbf{x}}^L(X)\right)\\ &\leq (L+1)^{|\mathcal{X}|}\sup_{P_\mathbf{x}:D(P_\mathbf{x}||P_S)>E_{2|1}^I}\exp\{-LD(P_\mathbf{x}||P_S)\}\\ &\leq \exp\left\{-L\left[E_{2|1} - o_L(1)\right]\right\},\end{aligned}$$

where $o_L(1) \to 0$ when $L \to \infty$. From the definition of the reliability we get the claim 3.

Now let us prove (6), for which the first we need to estimate second error probability $\alpha_{1|2}(\varphi_L^*)$.

Using (1), (2) and (14) for this case we obtain:

$$\begin{aligned}\alpha_{1|2}(\varphi_L^*) &= P_C^L\left(\mathcal{B}_1^L\right)\\ &= P_C^L\left(\bigcup_{P_\mathbf{x}:D(P_\mathbf{x}||P_S)\leq E_{2|1}^I} \mathcal{T}_{P_\mathbf{x}}^L(X)\right)\\ &\leq (L+1)^{|\mathcal{X}|}\sup_{P_\mathbf{x}:D(P_\mathbf{x}||P_S)\leq E_{2|1}^I} P_C^L\left(\mathcal{T}_{P_\mathbf{x}}^L(X)\right) \qquad (15)\\ &\leq (L+1)^{|\mathcal{X}|}\sup_{P_\mathbf{x}:D(P_\mathbf{x}||P_S)\leq E_{2|1}^I}\exp\{-LD(P_\mathbf{x}||P_C)\}\end{aligned}$$

$$= \exp\left\{-L\left(\inf_{P_{\mathbf{x}}:D(P_{\mathbf{x}}||P_S)\leq E^I_{2|1}} D(P_{\mathbf{x}}||P_C) - o_L(1)\right)\right\}.$$

Moreover, we can prove the inverse inequality:

$$\begin{aligned}\alpha_{1|2}(\varphi^*_L) &= P^L_C\left(\mathcal{B}^L_1\right)\\ &= P^L_C\left(\bigcup_{P_{\mathbf{x}}:D(P_{\mathbf{x}}||P_S)\leq E^I_{2|1}} \mathcal{T}^L_{P_{\mathbf{x}}}(X)\right)\\ &\geq \sup_{P_{\mathbf{x}}:D(P_{\mathbf{x}}||P_S)\leq E^I_{2|1}} P^L_C(\mathcal{T}^L_{P_{\mathbf{x}}}(X)) \qquad (16)\\ &\geq (L+1)^{-|\mathcal{X}|} \sup_{P_{\mathbf{x}}:D(P_{\mathbf{x}}||P_S)\leq E^I_{2|1}} \exp\{-LD(P_{\mathbf{x}}||P_C)\}\\ &= \exp\left\{-L\left(\inf_{P_{\mathbf{x}}:D(P_{\mathbf{x}}||P_S)\leq E^I_{2|1}} D(P_{\mathbf{x}}||P_C) + o_L(1)\right)\right\}.\end{aligned}$$

Taking into account (15), (16), and the continuity of the functions $D(P_{\mathbf{x}}||P_C)$ and $D(P_{\mathbf{x}}||P_S)$ we get that $\lim\limits_{L\to\infty} -L^{-1}\log\alpha_{1|2}(\varphi^*_L)$ exists:

$$\lim_{L\to\infty} -\frac{1}{L}\log\alpha_{1|2}(\varphi^*_L) = \inf_{P:D(P||P_S)\leq E^I_{2|1}} D(P||P_C).$$

On the other hand,

$$\lim_{L\to\infty} -\frac{1}{L}\log\alpha_{1|2}(\varphi^*_L) = E^I_{1|2}(\varphi^*) \overset{\triangle}{=} E^{*,I}_{1|2}.$$

This means that for given $E^I_{1|2}$ there exists φ^* test, for which

$$E^{*,I}_{1|2} = E^{*,I}_{1|2}(E^I_{2|1}) = \inf_{P:D(P||P_S)\leq E^I_{2|1}} D(P||P_C).$$

The proof of the first part of Theorem 1 will be accomplished if we demonstrate that the sequence of the test φ^* is LAO, that is for given $E^I_{2|1}$ and every sequence of tests φ $E^I_{1|2}(\varphi) \leq E^{*,I}_{1|2}$ takes place.

Let us consider any other sequence φ^{**} of tests which for given $E^I_{2|1}$ is defined by partition of $\mathcal{X}^L$ to disjoint subsets $\mathcal{D}^L_1$ and $\mathcal{D}^L_2$ such that $E_{1|2}(\varphi^{**}) \geq E^{*,I}_{1|2}$. This condition is equivalent to the inequality

$$\alpha_{1|2}(\varphi^{**}_L) \leq \alpha_{1|2}(\varphi^*_L) \qquad (17)$$

for L large enough.

Let us show that $\mathcal{D}_2^L \bigcap \mathcal{B}_1^L = \emptyset$. If $\mathcal{D}_2^L \bigcap \mathcal{B}_1^L \neq \emptyset$, then there exists $P'_{\mathbf{x}}$ such that $D(P'_{\mathbf{x}}||P_S) \leq E^I_{2|1}$ and $\mathcal{T}^L_{P'_{\mathbf{x}}}(X) \in \mathcal{D}_2^L$ from which it follows that

$$\alpha_{2|1}(\varphi_L^{**}) = P_S^N(\mathcal{D}_2^L) \geq P_S^L(\mathcal{T}^N_{P'_{\mathbf{x}}}(X)) \geq \exp\left\{-L\left[E^I_{2|1} + o_L(1)\right]\right\}.$$

From $\mathcal{D}_1^L \bigcup \mathcal{D}_2^L = \mathcal{X}^L$, $\mathcal{D}_1^L \bigcap \mathcal{D}_2^L = \emptyset$ and $\mathcal{D}_2^L \bigcap \mathcal{B}_1^L = \emptyset$, follows that $\mathcal{B}_1^L \subseteq \mathcal{D}_1^L$. If $\mathcal{B}_1^L \subset \mathcal{D}_1^L$, then we have that $\alpha_{1|2}(\varphi_L^{**}) \geq \alpha_{1|2}(\varphi_L^{*})$, which contradicts to (17). Hence $\mathcal{D}_1^L = B_1^L$, as well as $\mathcal{D}_2^L = B_2^L$. It is the same that $\varphi^{**} = \varphi^{*}$.

The proof of the second part of the Theorem 1 is simple. Really, if $E_{2|1} \geq D(P_C||P_S)$, then from (6) follows that $E^{*,I}_{1|2}$ is equal to 0.

The theorem is proved.

The Proof of Theorem 2.

The optimal division of the set $(\mathcal{M} \times \mathcal{K})^N$ is constructed in the following way.

$$\mathcal{B}_1^N = \bigcup_{Q_{\mathbf{m},\mathbf{k}}:\ D(Q_{\mathbf{m},\mathbf{k}}||P_{MK}) \leq E^{II}_{2|1}} \mathcal{T}^N_{Q_{\mathbf{m},\mathbf{k}}}(M,K),$$

$$\mathcal{B}_2^N = \bigcup_{Q_{\mathbf{m},\mathbf{k}}:\ D(Q_{\mathbf{m},\mathbf{k}}||P_S) > E^{II}_{2|1}} \mathcal{T}^N_{Q_{\mathbf{m},\mathbf{k}}}(M,K).$$

We must prove, that this division for given $E^{II}_{2|1}$ determines a test Φ^*_N.

The following equalities take place for $(\mathbf{m},\mathbf{k}) = [(m_1,k_1),(m_2,k_2),...,(m_N,k_N)]$, $(\mathbf{m},\mathbf{k}) \in \mathcal{T}^N_{Q_{\mathbf{m},\mathbf{k}}}(M,K)$

$$P_M^N P_K^N(\mathbf{m},\mathbf{k}) = \exp\Big\{-N\Big[D(Q_{\mathbf{m},\mathbf{k}}||P_M P_K) + H_{Q_{\mathbf{m},\mathbf{k}}}(M,K)\Big]\Big\}, \qquad (18)$$

and

$$P_{MK}^N(\mathbf{m},\mathbf{k}) = \exp\Big\{-N\Big[D(Q_{\mathbf{m},\mathbf{k}}||P_{MK}) + H_{Q_{\mathbf{m},\mathbf{k}}}(M,K)\Big]\Big\}. \qquad (19)$$

Using (2), (4) and (19) we will get the estimation of the first kind error probability $\alpha^I_{2|1}(\varphi^*_L)$:

$$\alpha_{2|1}(\Phi^*_N) = P^N_{MK}\left(\mathcal{B}_2^N\right)$$

$$= P^N_{MK}\left(\bigcup_{Q_{\mathbf{m},\mathbf{k}}:D(Q_{\mathbf{m},\mathbf{k}}||P_{MK}) > E^{II}_{2|1}} \mathcal{T}^N_{Q_{\mathbf{m},\mathbf{k}}}(M,K)\right)$$

$$\leq (N+1)^{|\mathcal{M}||\mathcal{K}|} \sup_{Q_{\mathbf{m},\mathbf{k}}:D(Q_{\mathbf{m},\mathbf{k}}||P_{MK})>E^{II}_{2|1}} P^N_{MK}\left(\mathcal{T}^N_{Q_{\mathbf{m},\mathbf{k}}}(M,K)\right)$$

$$= (N+1)^{|\mathcal{M}||\mathcal{K}|} \sup_{Q_{\mathbf{m},\mathbf{k}}:D(Q_{\mathbf{m},\mathbf{k}}||P_{MK})>E^{II}_{2|1}} \exp\Big\{ - ND(Q_{\mathbf{m},\mathbf{k}}||P_{MK})\Big\}$$

$$\leq \exp\left\{-N\left[E^{II}_{2|1} - o_N(1)\right]\right\},$$

where $o_N(1) \rightarrow 0$ when $N \rightarrow \infty$.

Now let us consider the second kind error probability. Using (2), (4) and (18) we can find the lower and the upper bounds of the second kind error probability $\alpha_{1|2}(\Phi^*_N)$.

The following inequality gives the upper bound:

$$\begin{aligned}\alpha_{1|2}(\Phi^*_N) &= P^N_M P^N_K\left(\mathcal{B}^N_1\right)\\ &= P_M P_K\left(\bigcup_{Q_{\mathbf{m},\mathbf{k}}:D(Q_{\mathbf{m},\mathbf{k}}||P_{MK})\leq E^{II}_{2|1}} \mathcal{T}^N_{Q_{\mathbf{m},\mathbf{k}}}(M,K)\right)\\ &\leq (N+1)^{|\mathcal{M}||\mathcal{K}|} \sup_{Q_{\mathbf{m},\mathbf{k}}:D(Q_{\mathbf{m},\mathbf{k}}||P_{MK})\leq E^{II}_{2|1}} P_M P_K\left(\mathcal{T}^N_{Q_{\mathbf{m},\mathbf{k}}}(M,K)\right) \qquad (20)\\ &\leq \exp\left\{-N\left[\inf_{Q_{\mathbf{m},\mathbf{k}}:\ D(Q_{\mathbf{m},\mathbf{k}}||P_{MK})\leq E^{II}_{2|1}} D(Q_{\mathbf{m},\mathbf{k}}||P_M P_K) - o_N(1)\right]\right\}.\end{aligned}$$

The lower bound is:

$$\begin{aligned}\alpha_{1|2}(\Phi^*_N) &= P^N_{MK}\left(\mathcal{B}^N_1\right)\\ &= P_M P_K\left(\bigcup_{Q_{\mathbf{m},\mathbf{k}}:D(Q_{\mathbf{m},\mathbf{k}}||P_{MK})\leq E^{II}_{2|1}} \mathcal{T}^N_{Q_{\mathbf{m},\mathbf{k}}}(M,K)\right)\\ &\geq \sup_{Q_{\mathbf{m},\mathbf{k}}:D(Q_{\mathbf{m},\mathbf{k}}||P_{MK})\leq E^{II}_{2|1}} P_M P_K\left(\mathcal{T}^N_{Q_{\mathbf{m},\mathbf{k}}}(M,K)\right) \qquad (21)\\ &\geq (N+1)^{-|\mathcal{M}||\mathcal{K}|} \sup_{Q_{\mathbf{m},\mathbf{k}}:D(Q_{\mathbf{m},\mathbf{k}}||P_{MK})\leq E_{2|1}} \exp\{-ND(Q_{\mathbf{m},\mathbf{k}}||P_{MK})\\ &\geq \exp\left\{-N\left[\inf_{Q_{\mathbf{m},\mathbf{k}}:\ D(Q_{\mathbf{m},\mathbf{k}}||P_{MK})\leq E^{II}_{2|1}} D(Q_{\mathbf{m},\mathbf{k}}||P_M P_K) + o_N(1)\right]\right\}.\end{aligned}$$

According to (20), (21), the definition of reliability and the continuity of the divergence function we get

$$E^{II}_{1|2}(\Phi^*_N) \overset{\triangle}{=} E^{*,II}_{1|2} = \inf_{Q:\ D(Q||P_{MK}) \leq E^{II}_{2|1}} D(Q||P_M P_K).$$

The proof of the optimality of the test Φ^* is similar to the proof of Theorem 1. The theorem is proved.

Acknowledgements

The work was supported by the Science Committee of RA, in the frames of the research project № 21T-1B151.

References

[Alomair and Poovendran 2009] Alomair, B, Poovendran, R: "Information theoretically secure encryption with almost free authentication"; JUCS - Journal of Universal Computer Science, 15, 15 (2009), 2937-2956. https://doi.org/10.3217/jucs-015-15-2937

[Augot et al. 2011] Augot, D., Barbier, M., Fontaine, C.: "Ensuring message embedding in wet paper steganography"; Cryptography and Coding, IMACC 2011, Lecture Notes in Computer Science, 7089, Springer, Berlin, Heidelberg, (2011), 244-258. https://doi.org/10.1007/978-3-642-25516-8_15

[Backes and Cachin 2005] Backes, M., Cachin, C.: "Public-key steganography with active attacks"; Lecture Notes in Computer Science, 3378, Springer-Verlag, (2005), 210-226. https://doi.org/10.1007/978-3-540-30576-7_12

[Balado and Haughton 2018] Balado, F., Haughton, D.: "Asymptotically optimum perfect universal steganography of finite memoryless sources"; IEEE Transactions on Information Theory, 64, 2 (2018), 1199-1216. https://doi.org/10.1109/TIT.2017.2783539

[Blahut 1974] Blahut, R,: "Hypothesis testing and information theory"; IEEE Transactions on Information Theory, 20, 4 (1974), 405–417. https://doi.org/10.1109/TIT.1974.1055254

[Blahut 1987] Blahut, R.: "Principles and Practice of Information Theory"; Addison-Wesley, Reading, MA (1987).

[Birgé 1981] Birgé, L.: "Vitesses maximales de décroissance des erreurs et tests optimaux associés"; Z. Wahrsch. verw. Gebiete, 55, (1981), 261–273.

[Cachin et al. 1998] Cachin, C.: "An information-theoretic model for steganography"; Proc. Workshop on Information Hiding, Lecture Notes in Computer Science, 1525, Springer-Verlag, (1998), 306-318. https://doi.org/10.1007/3-540-49380-8_21

[Cachin 2004] Cachin, C.: "An information-theoretic model for steganography"; Information and Computation, 192, (2004), 41–56. https://doi.org/10.1016/j.ic.2004.02.003

[Csiszár 1998] Csiszár, I.: "Method of types"; IEEE Transactions on Information Theory, 44, 6 (1998), 2505-–2523. https://doi.org/10.1109/18.720546

[Csiszár and Körner 1981] Csiszár, I., Körner, J.: "Information Theory: Coding Theorems for Discrete Memoryless Systems"; Academic Press, New York (1981).

[Csiszár and Longo 1971] Csiszár, I., Longo, G.: "On the error exponent for source coding and for testing simple statistical hypotheses"; Studia Sc. Math. Hungarica, 6, (1971), 181–191.

[Cover and Thomas 2006] Cover, T., Thomas, J.: "Elements of Information Theory"; Second Edition, Wiley, New York (2006).

[Dedic et al. 2009] Dedic, N., Itkis, G., Reyzin, L., Russell, S.: “Upper and lower bounds on black box steganography”; Journal of Cryptology, 22, 3 (2009), 365-394. https://doi.org/10.1007/s00145-008-9020-3

[Grigoryan and Harutyunyan 2015] Grigoryan, N., Harutyunyan, A.: “Multiple hypothesis testing for arbitrarily varying sources”; Communications in Information and Systems, 15, 3 (2015), 309–330. https://doi.org/10.4310/CIS.2015.v15.n3.a1

[Grigoryan et al. 2011] Grigoryan, N., Harutyunyan, A., Voloshynovskiy, S., Koval, O.: “On multiple hypothesis testing with rejection option”; Proc. IEEE Information Theory Workshop, Paraty, Brazil, (Oct 2011). doi: 10.1109/ITW.2011.6089531.

[Haroutunian 1989] Haroutunian, E.: “On asymptotically optimal criteria for Markov chains”; The First World Congress of Bernoulli Society, (in Russian) 2, 3 (1989), 153–156.

[Haroutunian 1990] Haroutunian, E.: “Logarithmically asymptotically optimal testing of multiple statistical hypotheses”; Problems of Control and Information Theory, 19, 5-6 (1990), 413–421.

[Harutyunyan et al. 2011] Harutyunyan, A., Grigoryan, N., Voloshynovskiy, S., Koval, O.: “A new biometric identification model and the multiple hypothesis testing for arbitrarily varying objects”; Proc. Special Interest Group on Biometrics and Electronic Signatures (BIOSIG2011), Darmstadt, Germany, (Sep 2011), 305–312.

[Haroutunian et al. 2008] Haroutunian, E., Haroutunian, M., Harutyunyan, A.:“Reliability Criteria in Information Theory and in Statistical Hypothesis Testing”; Foundations and Trends in Communications and Information Theory, 4, 2-3 (2008). doi: 10.1561/0100000008

[Haroutunian et al. 2022] Haroutunian, M., Hakobyan, P., Harutyunyan, A., Avetisyan, A.: “Information-theoretic investigation of authenticated steganographic Model in the presence of active adversary”; CODASSCA Intern. Workshop, (2022), 1-7. https://doi.org/10.30819/5520

[Haroutunian et al. 2018] Haroutunian, M., Haroutunian, E., Hakobyan, P., Mikayelyan, H.: “Logarithmically asymptotically optimal testing of statistical hypotheses in steganography applications”; CODASSCA Intern. Workshop, (2018), 157–163.

[Harutyunyan and Han Vinck 2006] Harutyunyan, A., Han Vinck, A. J.: “Error exponent in AVS coding”; Proc. IEEE International Symposium on Information Theory, Seattle, WA, 2 (July 2006), 166-2170. doi:10.1109/ISIT.2006.261934

[Hoeffding 1965] Hoeffding, W.: “Asymptotically optimal tests for multinomial distributions”; The Annals of Mathematical Statistics, 36, (1965), 369–401. doi: 10.1214/aoms/1177700150

[Hopper et al. 2002] Hopper, N., Langford, J., von Ahn, L.: “Provably secure steganography”; Lecture Notes in Computer Science, 2442, Springer-Verlag, (2002), 18-22. https://doi.org/10.1007/3-540-45708-9_6

[Katzenbeisser et al. 2002] Katzenbeisser, S., Petitcolas, F. A. P.: “Defining security in steganographic systems”; Security and Watermarking of Multimedia Contents IV, 4675, (2002), 50-56. https://doi.org/10.1117/12.465313

[Liśkiewicz et al. 2011] Liśkiewicz, M., Reischuk, R., Wölfel, U.: “Grey-box steganography”; Theory and Applications of Models of Computation. TAMC 2011. Lecture Notes in Computer Science, 6648, Springer, Berlin, Heidelberg (2011), 390-402. https://doi.org/10.1007/978-3-642-20877-5_38

[Longo and Sgarro 1980] Longo, G., Sgarro, A.: “The error exponent for the testing of simple statistical hypotheses: A combinatorial approach”; Journal of Combinatorics, Information and System Sciences, 5, 1 (1980), 58-67.

[Maurer 2000] Maurer, U.: “Authentication theory and hypothesis testing”; IEEE Transactions on Information Theory, 46, 4 (2000), 1350–1356. doi: 10.1109/18.850674.

[Mittelholzer 2000] Mittelholzer, T.: “An information-theoretic approach to steganography and watermarking”; Proc. Workshop on Information Hiding, Lecture Notes in Computer Science, 1768, Springer-Verlag (2000), 1-16. https://doi.org/10.1007/10719724_1

[Moulin and O’Sullivan 2003] Moulin, P., O’Sullivan, J. A.: “Information theoretic analysis of information hiding”; IEEE Transactions on Information Theory, 49, 3 (2003), 563-593. doi: 10.1109/TIT.2002.808134.

[O’Sullivan et al. 1998] O’Sullivan, J. A., Moulin, P., Ettinger, J. M.: “Information theoretic analysis of steganography”; Proc. IEEE Intern. Symposium on Information Theory, 297 (1998). doi: 10.1109/ISIT.1998.708902.

[Pfitzmann 1996] Pfitzmann, B.: “Information hiding terminology”; Information Hiding, First International Workshop, Lecture Notes in Computer Science 1174, Springer (1996), 347–350. https://doi.org/10.1007/3-540-61996-8_52

[Shikata and Matsumoto 2008] Shikata, J., Matsumoto, T.: “Unconditionally secure steganography against active attacks”; IEEE Transactions on Information Theory, 54, 6 (June 2008), 2690–2705. doi: 10.1109/TIT.2008.921884.

[Tusnady 1977] Tusnady, G.: “On asymptotically optimal tests”; Ann. of Statist, 5, 2 (1977), 385-393. https://doi.org/10.1214/aos/1176343804

[Von Ahn and Hopper 2004] Von Ahn, L., Hopper, N.: “Public-key steganography”; Lecture Notes in Computer Science 3027, Springer-Verlag (2004), 323-341. https://doi.org/10.1007/978-3-540-24676-3_20

[Wang and Moulin 2008] Wang, Y., Moulin, P.: “Perfectly secure steganography: capacity, error exponents, and code constructions”; IEEE Transactions on Information Theory, 54, 6 (2008), 2706-2722. https://doi.org/10.1109/TIT.2008.921684

[Willems et al 2003] Willems, F., Kalker, T., Goseling, J., and Linnartz, J.-P.: “On the capacity of a biometrical identification system”; Proc. IEEE International Symposium on Information Theory, Yokohama, Japan (July 2003). doi: 10.1109/ISIT.2003.1228096.

[Yagi and Hirasawa 2022] Yagi, H., Hirasawa, Sh.: “Performance analysis for biometric identification systems with nonlegitimate users”; Proc. 2022 IEEE International Conference on Systems, Man, and Cybernetics (SMC), Prague, Check Republic (Oct 2022), 3060-3065. doi: 10.1109/SMC53654.2022.9945401.

Journal of Universal Computer Science, vol. 29, no. 11 (2023), 1274-1297
submitted: 19/5/2023, accepted: 30/9/2023, appeared: 28/11/2023 CC BY-ND 4.0

Distributed Tracing for Troubleshooting of Native Cloud Applications via Rule-Induction Systems

Arnak Poghosyan
(Institute of Mathematics of NAS RA, Yerevan, Armenia
VMware, Palo Alto, US
American University of Armenia, Yerevan, Armenia
https://orcid.org/0000-0002-6037-4851
arnak@instmath.sci.am, apoghosyan@vmware.com, apoghosyan@aua.am)

Ashot Harutyunyan
(AI Lab at Yerevan State University, Yerevan, Armenia
Institute for Informatics and Automation Problems of NAS RA, Yerevan, Armenia
VMware, Palo Alto, US
https://orcid.org/0000-0003-2707-1039, aharutyunyan@vmware.com)

Naira Grigoryan
(VMware, Palo Alto, US
https://orcid.org/0000-0003-3980-4500, ngrigoryan@vmware.com)

Clement Pang
(VMware, Palo Alto, US
https://orcid.org/0000-0002-5821-0735, clementp@gmail.com)

Abstract: Diagnosing IT issues is a challenging problem for large-scale distributed cloud environments due to complex and non-deterministic interrelations between the system components. Modern monitoring tools rely on AI-empowered data analytics for detection, root cause analysis, and rapid resolution of performance degradation. However, the successful adoption of AI solutions is anchored on trust. System administrators will not unthinkingly follow the recommendations without sufficient interpretability of solutions. Explainable AI is gaining popularity by enabling improved confidence and trust in intelligent solutions. For many industrial applications, explainable models with moderate accuracy are preferable to highly precise black-box ones. This paper shows the benefits of rule-induction classification methods, particularly RIPPER, for the root cause analysis of performance degradations. RIPPER reveals the causes of problems in a set of rules system administrators can use in remediation processes. Native cloud applications are based on the microservices architecture to consume the benefits of distributed computing. Monitoring such applications can be accomplished via distributed tracing, which inspects the passage of requests through different microservices. We discuss the application of rule-learning approaches to trace traffic passing through a malfunctioning microservice for the explanations of the problem. Experiments performed on datasets from cloud environments proved the applicability of such approaches and unveiled the benefits.

Keywords: cloud-native applications, application troubleshooting, distributed tracing, RED metrics, root cause analysis, explainable AI, rule-induction systems, RIPPER
Categories: I.2.6, I.5.4
DOI: 10.3897/jucs.112513

1 Introduction

Identification and remediation of the performance degradations of cloud applications require automated, real-time, and intelligent root cause analysis (RCA). Due to the complexity of microservices architecture (see [Knoche and Hasselbring, 2019, Lin et al., 2018, Cai et al., 2019, Liu et al., 2021, Tzanettis et al., 2022] with references therein), administrators are unable to perform timely detection and identification of IT issues. ML/AI empowered data analytics, or AI Ops (see [Notaro et al., 2020]), is designed to identify and resolve service incidents by supporting domain experts. RCA is one of the critical capabilities of AI Ops that will explain IT incidents in a human-readable medium through highly interpretable ML models. We show how rule-learning systems like RIPPER (see [Cohen, 1995]) can be the foundation of explainable RCA.

Application performance monitoring (APM) tools (see [Heger et al., 2017]) collect all available information for effectively addressing performance issues and managing applications (see [Harutyunyan et al., 2022, Harutyunyan et al., 2020a, Harutyunyan et al., 2020b, Poghosyan et al., 2016, Harutyunyan et al., 2018, Mahmud et al., 2021, Elsaadawy et al., 2019, Klaise et al., 2020, Vitali, 2022]). One of the main goals of APM is to enable the observability of cloud applications by aggregating data and outlining the system's overall health. Time series data, log data, and traces are the pillars of observability. This paper shows how explainable RCA can be built on top of application traces via rule-induction methods.

Distributed tracing (see [Heger et al., 2017, Parker et al., 2020] with references therein) is a modern technology for monitoring native cloud applications with microservices architecture. It is one of the best-known approaches for the monitoring of distributed systems. Traces observe end-to-end requests propagating through the distributed microservices and detect transaction slowdowns (see [VMware, 2022]). A single trace shows an individual request passage through the microservices. It contains a series of tagged time intervals known as spans (see Figure 1).

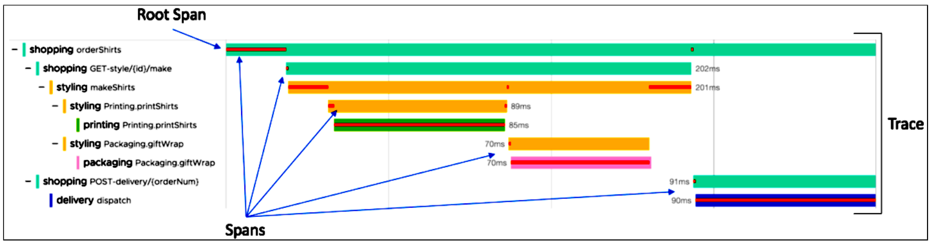
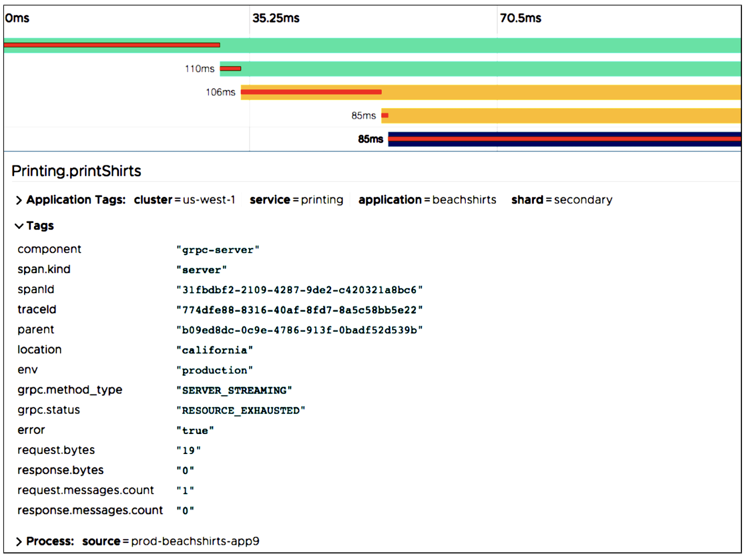

Root Span
shopping orderShirts
shopping GET-style/{id}/make
styling makeShirts
styling Printing.printShirts
printing Printing.printShirts
styling Packaging.giftWrap
packaging Packaging.giftWrap
shopping POST-delivery/{orderNum}
delivery dispatch
202ms
201ms
89ms
85ms
70ms
70ms
91ms
90ms
Trace
Spans

Figure 1: Individual trace of a simulated application (https://docs.wavefront.com/trace_data_details.html).

A span contains metadata known as tags and application tags for better process resolution. Figure 2 shows the tags for the span ”printShirts” from the service ”printing.” It shows the tags ”cluster,” ”service,” ”location,” ”parent,” ”env”, etc. One of the essential fields is the tag ”error,” which can be used to label traces. This field's value=”true” indicates that the corresponding trace is erroneous. Otherwise, a trace is expected if the field is missing (typical). We can also label traces based on other characteristics.

The troubleshooting of an application in case of some issues can be manually performed via traces browser (see Figure 3), which shows a group of traces corresponding



Printing.printShirts

> Application Tags: **cluster** = us-west-1 **service** = printing **application** = beachshirts **shard** = secondary

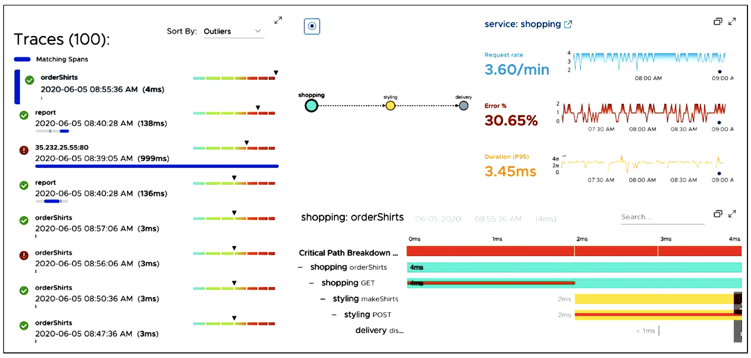
∨ Tags

component	"grpc-server"
span.kind	"server"
spanId	"31fbdbf2-2109-4287-9de2-c420321a8bc6"
traceId	"774dfe88-8316-40af-8fd7-8a5c58bb5e22"
parent	"b09ed8dc-0c9e-4786-913f-0badf52d539b"
location	"california"
env	"production"
grpc.method_type	"SERVER_STREAMING"
grpc.status	"RESOURCE_EXHAUSTED"
error	"true"
request.bytes	"19"
response.bytes	"0"
request.messages.count	"1"
response.messages.count	"0"

> Process: **source** = prod-beachshirts-app9

Figure 2: The tags of a span of a simulated application (https://docs.wavefront.com/tracing_traces_browser.html).

to a service, unveils their durations, indicates anomaly traces, and for each trace presents the structure of spans.


Traces (100):

Sort By: Outliers

Matching Spans

orderShirts
2020-06-05 08:55:36 AM (4ms)

report
2020-06-05 08:40:28 AM (138ms)

35.232.25.55:80
2020-06-05 08:39:05 AM (999ms)

report
2020-06-05 08:40:28 AM (136ms)

orderShirts
2020-06-05 08:57:06 AM (3ms)

orderShirts
2020-06-05 08:56:06 AM (3ms)

orderShirts
2020-06-05 08:50:36 AM (3ms)

orderShirts
2020-06-05 08:47:36 AM (3ms)

service: shopping

Request rate 3.60/min

Error % 30.65%

Duration (P95) 3.45ms

shopping: orderShirts 06-05-2020 08:55:36 AM (4ms)

Figure 3: Traces browser for a simulated application (https://docs.wavefront.com/tracing_traces_browser.html).

The health of a service can be tracked by the values of RED (rate, error, duration) metrics corresponding to the number of requests per minute, the number of errors (failed requests per minute), and the p95 quantile of trace durations in a minute. The change in behaviors of RED metrics can indicate some issues.

Moreover, some hard thresholds can be put on top of those metrics, which violations will trigger alerts indicating users to start the process of troubleshooting for the root cause analysis of issues. Unfortunately, all those tasks can be performed only for a limited number of services and traces. Universal manual troubleshooting for all available services and traces is not feasible due to the complexity of interrelations between application components and a large volume of trace information. We consider a more general and automated approach to the problem of application troubleshooting based on trace traffic.

Trace traffic shows how applications and services interact. Application map is the visualization of the tracing traffic (see Figure 4). The colors of services indicate the statuses of the corresponding microservices. The colors optionally can be assigned by the values of RED metrics. The red-colored microservices have poor performances, and our main goal is to understand/explain those problems. We can select a poor-behaving microservice, collect tracing traffic passing through it, and analyze via explainable AI (XAI) methods for some acceptable interpretations of the performance degradations in trace types, spans, and tags.

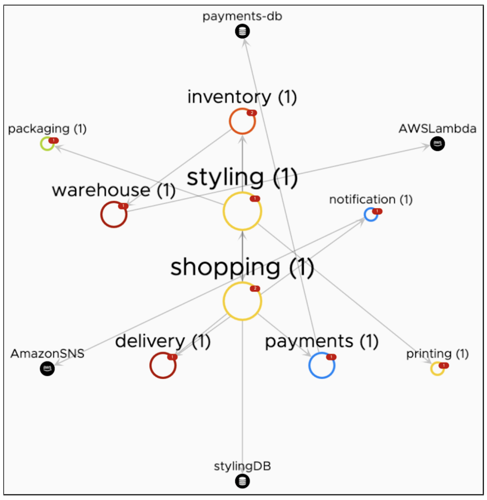


Figure 4: Application map for a simulated application (https://docs.wavefront.com/tracing_ui_overview.html).

2 Related Work

System administrators can no longer perform real-time decision-making due to the growth of large-scale distributed cloud environments with complicated, invisible underlying processes. Those systems require more advanced and ML/AI-empowered intelligent RCA with explainable and actionable recommendations (see [Solé et al., 2017, Harutyunyan et al., 2019, Poghosyan et al., 2021a, Poghosyan et al., 2016, Poghosyan et al., 2021b, Marvasti et al., 2013] with references therein). XAI (see [Barredo Arrieta et al., 2020]) builds user and AI trust, increases solutions' satisfaction, and leads to more actionable and robust prediction and root cause analysis models. Many users think it is risky to trust and follow AI recommendations and predictions blindly, and they need to understand the foundation of those insights. Many ML approaches like decision trees and rule-induction systems (see [Fürnkranz et al., 2012]) have sufficient explainability capabilities for RCA. They can detect and predict performance degradations and identify the most critical features (processes) potentially responsible for the malfunctioning. In many applications, explainable outcomes can be more valuable than conclusions based on more powerful approaches that act like black boxes. Rule learners are the best if outcomes' simplicity and human interpretability are superior to the predictive power (see [Fürnkranz et al., 2012]). The list of known rule learners consists of many exciting approaches like association rules, APRIORI, CN2, AQ, FOIL, STUCCO, OPUS, RIPPER, and many others. We refer to [Fürnkranz et al., 2012] for a more detailed description of available algorithms, their comparisons, and historical analysis. It contains relatively rich references and describes several applications.

Data overfitting is a well-known issue that also affects rule learning systems. RIPPER (see [Cohen, 1995]) was the first rule-learning system that effectively countered the overfitting problem (see [Fürnkranz and Kliegr, 2015, Fürnkranz, 1997]). It is based on several previous works, most notably the incrementally reduced error pruning (IREP) idea described in [Fürnkranz and Widmer, 1994]. In several independent studies, RIPPER has proved to be among the most competitive rule-learning algorithms available today. It is competitive with C4.5RULES (see [Quinlan, 2014]) without losing IREP's efficiency. RIPPER is a classification rule induction approach, and the crucial steps for its realization are data labeling and feature selection/extraction in the case of massive datasets. Many authors have tried to improve it. An interesting approach is FURIA (see [Hühn and Hüllermeier, 2009]).

Applying rule learning algorithms to industrial and IT problems has a long history. Papers [Suriadi et al., 2013, Lin et al., 2020] consider analysis based on log data. Paper [Suriadi et al., 2013] proposes an approach to enrich and transform process-based logs for RCA by applying classification techniques. The idea is to transform event logs by using classical data mining techniques. The final classification-ready problem was successfully treated by J48 and JRip, available in the machine learning library Weka. Algorithm J48 is the Weka implementation of the well-known C4.5RULES learning algorithm (see [Quinlan, 2014]). Its improvement is known as algorithm C5.0. Algorithm JRip is the Weka implementation of RIPPER. It resulted in much simpler rules with comparable accuracy of classification. Paper [Lin et al., 2020] considers the problem of RCA in a large-scale production environment based on structured logs. The challenge is the complexity of services running across global data centers. The authors explore the application of the Apriori algorithm (see [Agrawal et al., 1993]) with subsequent improvement by the FP-Growth approach (see [Han et al., 2004]).

Papers [Lee and Stolfo, 1998, Helmer et al., 1998, Helmer et al., 2002] consider the problem of intrusion and abuse detection in computer systems in networked environments. Their method is based on the system calls executed by a program like primitive traces. Paper [Lee and Stolfo, 1998] experimented with "sendmail" system call and network "tcpdump" data. In the first example, each trace data file has two columns of integers. The first is the process IDs, and the second is the system call names. The set includes known normal and abnormal traces. Then, they apply RIPPER for a rule induction in a specific manner - each record has the same number of positional attributes plus a label. The final rules were used to predict whether a sequence is abnormal or normal. Experiments showed that the expected behavior of a program execution could be established and used to detect its anomaly states. RIPPER was also applied to "tcpdump" data for anomaly predictions. Furthermore, the authors combined different classifiers to improve the effectiveness of detecting intrusions. The list included association rules and frequent episodes algorithm (see [Mannila et al., 1995]) with promising results.

Paper [Helmer et al., 1998] continued the previous work by proposing a feature vector construction technique for system call traces. The traces were encoded as binary-valued bits in feature vectors. Each bit in the vector indicates whether a known system call sequence appeared during the execution of a process. Those feature vectors were used for rule induction via RIPPER.

Paper [Helmer et al., 2002] went further and discussed an interesting approach for complexity reduction of learning algorithms. Naturally, the complexity is directly connected with the number of features involved in the learning. It should be a very reasonable application of learning systems to important features. Feature subset selection has been shown to improve the performance of a learning algorithm and reduce the effort and amount of data required for machine learning on a broad range of problems (see [Liu and Motoda, 1998]). Paper [Helmer et al., 2002] considered genetic algorithms for feature subset selection (see also [John et al., 1994]). The main open issue of those papers is the application of the methods to heterogeneous distributed systems. The current paper addresses this problem and describes a far more general approach to feature construction and data labeling with different levels of resolution.

The current paper is the extension of our paper published in the 3rd CODASSCA Workshop on Collaborative Technologies and Data Science in Artificial Intelligence Applications, Yerevan 2022, Armenia (see [Poghosyan et al., 2022]). Several US patents are available regarding the ideas of this paper (see [Nag et al., 2021, Poghosyan et al., 2021c]).

3 Methodology

RCA's main goal is to understand a problem's root causes to identify appropriate resolution procedures. Even if the root causes are not directly visible, RCA must provide sufficient explainability of a problem for the acceleration of its remediation or to prevent future occurrences. It means that the capabilities of detection or prediction of performance degradations need to be improved for ML-empowered RCA solutions. A sufficient level of explainability is the main requirement from those ML solutions for spreading light onto the underlying complex processes.

Fortunately, there are many powerful learning approaches with a high level of explainability, like decision trees and rule-induction methods. In this paper, we illustrate the power of RIPPER (see [Cohen, 1995]) to explain the performance degradations of application microservices monitored via distributed tracing. According to [Fürnkranz and Kliegr, 2015], RIPPER is still state-of-the-art in inductive rule learning. It has some important technical characteristics, like supporting missing values, numerical and categorical variables, and multiple classes. According to [Cohen, 1995], it scales nearly linearly with the number of instances in large datasets.

We experimented with Weka's RIPPER implementation, or JRip (see [Witten et al., 2005]). This implementation is very stable and fast. It takes seconds (up to a minute) to induce rules for a dataset with thousands of traces. The default algorithm is known as RIPPER2 (see [Cohen, 1995]). JRip has a more general implementation known as RIPPERk that repeatedly optimizes k times. We tried RIPPER5 and RIPPER10, which sometimes resulted in fewer rules but took longer for the execution.

RIPPER aims to find regularities in data in the form of an IF-THEN rule (see [Fürnkranz and Kliegr, 2015]). The condition of a rule (body) is composed of a conjunction of Boolean terms, each consisting of a constraint that needs to be satisfied by a trace. The rule is said to fire if all constraints are satisfied, and a trace is said to be covered by the rule. The rule head is a class label predicted in case the rule fires. The RIPPER rules can be simple (with only one condition) or complex (consisting of multiple constraints). The importance of a rule can be characterized by the following well-known measures (see [Fürnkranz and Kliegr, 2015]):

$$Confidence = N(body \to head)/N(body),$$

and

$$Coverage = N(body \to head)/N(positive\ class),$$

where $N(body \to head)$ is the number of traces that satisfy both the body and head of the rule (correct classifications by the corresponding rule), $N(body)$ is the number of traces that satisfy the body of the rule (both correct and incorrect classifications by the rule) and $N(positiveclass)$ is the number of traces labeled as “interesting for explanations.” We can set some thresholds for acceptable rules. Say, the coverage and confidence of the rules should be greater than 20% and 80%, respectively.

We recommended a slightly modified procedure for retrieving rules compared to the classical direct application of RIPPER. For many problems, several spans or tag values can equally explain the root cause. RIPPER will randomly select some of them for rule retrieval. However, not all acceptable rules will be equally relevant for system administrators for further remediation progress. We apply RIPPER in cycles (iterations) to show all hidden recommendations. This is somewhat similar to a feature extraction process but with extra caution, as we don’t know which attributes will lead to actionable recommendations without application domain knowledge. After each iteration of rule extraction, all features appearing in the previous rules should be removed from the dataset, and the rule-learning algorithm should be applied again. We iterate the procedure until some stopping criteria – time limitation or the weakness of the remaining rules.

4 Data Preparation and Algorithms

The analysis of tracing traffic passing through a specific microservice starts with data preprocessing. The traffic can contain hundreds or thousands of traces with different numbers of spans and tags, which are application-specific. Figure 5 shows a portion of trace traffic in a semi-structured format. It contains the names of traces as trace-IDs, then after each trace, the list of spans with process names, and after each span, the corresponding tags. Algorithm RIPPER requires tabular data (like dataframes in Pandas library). During the transformation, we remove some of the fields containing redundant information, useless for the insights of RCA, like ”traceID,” ”spanID,” ”startMs,” and many others. It should be very reasonable to denoise trace traffic based on expert knowledge or user feedback.

We consider three approaches for data tabularization, which lead to three algorithms with different levels of resolution and complexity. The first one (Algorithm A) works on a span level. We form the list of all distinct span names existing in trace traffic and put them as the names of the columns of a dataframe. Then, for each trace (as a row), we verify the names of its spans, and in the corresponding columns, enter value = 1. The other columns contain missing values as that specific trace doesn’t contain those spans. The entire dataframe consists of ones and missing values. Fortunately, RIPPER ignores those missing values and explains the output based on the existing spans in a trace. Eventually, the number of rows in the dataframe coincides with the number of

```
{'traceId': '6a538fb0-6a6f-4d5a-bbe7-9fd7295f939e',
 'spans': [{'name': 'ctp.auth.jdbc-execute',
   'host': 'a-atl1auth14',
   'startMs': 1597299097216,
   'durationMs': 1,
   'spanId': 'e8df55f9-648b-4d38-9e25-c634daf2c3e9',
   'traceId': '6a538fb0-6a6f-4d5a-bbe7-9fd7295f939e',
   'annotations': [{'parent': '76014304-3e43-47f5-a7be-b662a430afb0'},
    {'followsFrom': '76014304-3e43-47f5-a7be-b662a430afb0'},
    {'span.kind': 'client'},
    {'component': 'java-jdbc'},
    {'service': 'auth'},
    {'application': 'ctp'},
    {'shard': 'atl1'},
    {'cluster': 'agile'},
```

Figure 5: A portion of a trace-traffic in a semi-structured form.

traces in trace traffic, and the number of columns coincides with the number of distinct spans (processes).

Trace labeling can be performed via the tag ”error,” where the value=”true” indicates that the corresponding trace is erroneous. A missing tag ”error” indicates that the trace is normal. We can keep the number of erroneous traces slightly smaller than normal traces. In that case, RIPPER will assume that the class of erroneous traces is positive and will extract rules (containing only span names) for their explanations.

It can be useful also to construct a column in a dataframe that will indicate the type of a trace. Different ideas are known for trace-type determination. One of the ideas (see [Nag et al., 2021]) is to apply grouping of the traces based on the similarity of the sets of spans. Then, each group will define a trace type. However, technically, it is difficult to accomplish. The most straightforward idea is to use the root span available for some traces. All traces without root spans can be grouped as a single type. Hence, Algorithm A deals with spans and trace types to explain the set of erroneous traces. Sometimes, this level of resolution should be sufficient for problem identification.

The second approach (Algorithm B) works on a tag level. The process is similar to the previous one by using the list of span names in combination with tag names as the names of columns in the corresponding dataframe. Trace type and output columns can be copy-pasted from the previous construction. The dataframe cells contain value=1 for the corresponding spans, value=”NaN” for the missing ones, or the corresponding values of tags. This dataframe is much bigger than the first one, and Algorithm B is more complex than Algorithm A. RIPPER rules will contain the tags trying to explain the output by the values of tags. Hence, this approach has better resolution, and the corresponding insights may be more helpful.

The third approach (Algorithm C) uses a combined dataframe of the first two. This will be the most complex but the most complete approach. RIPPER rules will use both span names and tag names with their values, trying to find the best explanations.

RIPPER is relatively efficient in the sense of execution time. However, the number of spans and tags heavily affects the complexity. Those numbers are application-specific. In the case of hundreds of distinct spans, the execution of RIPPER can take several seconds up to a minute. This time range is connected with the noise in data. It is well-known that RIPPER is time-consuming for noisy datasets. In the case of thousands of distinct tags, the execution of RIPPER will take several minutes. The latest can be a problem in real-time monitoring systems when users are very demanding for fast executions. In those extreme situations, we can preprocess data for Algorithm A and show results regarding span names. Before a user can inspect the insights for remediation actions, we can proceed with Algorithms B or C for better problem explanations. In all cases, it is essential to measure the coverage and confidence of rules and limit the number of insights for users.

5 Experimental Results and Discussions

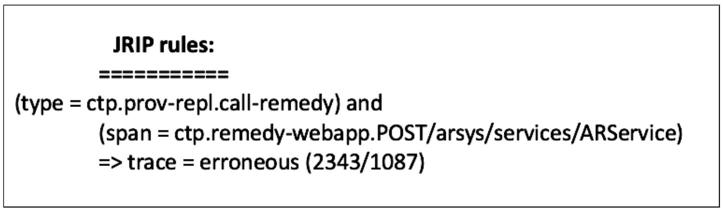
We perform experiments with data from actual cloud environments. Due to confidentiality, we will hide their names and use Customer I, II, and III. In all scenarios, we selected suspicious microservices from the corresponding application maps and collected some traces from the trace traffic passing through those components.

We aim to detect those spans, tags, and tag values that can explain the services' troubles. We show and discuss the outcomes of Algorithms A, B, and C.

5.1 Customer I

We selected 5428 traces for the first customer (Customer I). Algorithm A requires trace types and the distinct names of spans. Preliminary analysis showed 8 different trace types and 30 distinct spans. Hence, Algorithm A will work with the dataframe with 31 columns and 5428 rows. The first column contains information regarding the trace type. It is a categorical feature with 8 classes corresponding to different trace types. The remaining 30 columns contain values ”1” or ”NaN.” RIPPER ignores the missing values and explains the output based on the existing spans.

We performed data labeling based on erroneous traces. Our example contains 2393 erroneous traces and 3035 normal ones. We see that the number of erroneous traces is smaller than that of normal ones. RIPPER assumes that the smaller class is positive and tries to find the rules for explaining the erroneous traces. Algorithm A will solve that problem based on types and span names. Figure 6 shows the outcome of RIPPER applied to the described dataframe (Algorithm A).



```
        JRIP rules:
        ===========
(type = ctp.prov-repl.call-remedy) and
        (span = ctp.remedy-webapp.POST/arsys/services/ARService)
        => trace = erroneous (2343/1087)
```

Figure 6: The outcome of Algorithm A for the dataset of Customer I.

It shows a complex rule as the combination of two conditions. The first condition is the existence of type ”ctp.prov-repl.call-remedy”. The errors can be connected with the traces with the specified root span. The second condition shows the existence of the span ”ctp.remedy-webapp...ARService” in traces. In combination, it means that all traces with the specified root span and containing the span ”ctp.remedy-webapp...ARService” explain a portion of errors in the trace traffic.

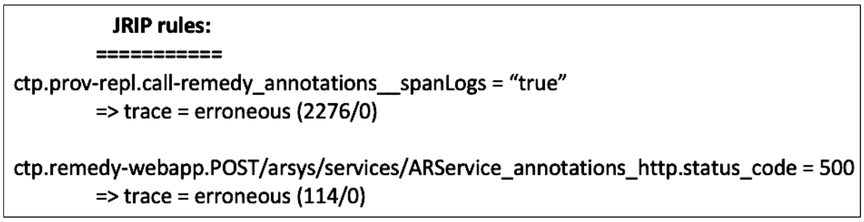
The fraction at the rule’s end will help calculate the corresponding scores. The numerator of the fraction 2343 shows how many times the rule has been fired. The denominator of the fraction 1087 shows the number of misclassifications. The rule has been fired for normal traces 1087 times.The coverage is 52% (see Table 1).

The Rules of Figure 6	Coverage	Confidence
Rule #1	0.52	0.54

Table 1: The coverage and confidence of the rules of Figure 6.

The confidence (the number of correct classifications divided by the number of fired rules) is 54% (see Table 1). The execution time of Algorithm A was less than a second. Unfortunately, the value of confidence is dissatisfactory. We will not show the rule to a user. We cannot explain the set of erroneous traces by the types or spans of traces with sufficient accuracy.

Let us continue with Algorithm B, which will explain the erroneous traces via tags. We found 489 distinct tags corresponding to different spans by storing them as span names plus tag names. The corresponding dataframe contains 489 columns and 5428 rows. The labels are the same. We can include or not include the column ”type.” Figure 7 shows the outcome of Algorithm B. We see two simple rules with perfect confidence 100%.



```
          JRIP rules:
        ===========
ctp.prov-repl.call-remedy_annotations__spanLogs = “true”
        => trace = erroneous (2276/0)

ctp.remedy-webapp.POST/arsys/services/ARService_annotations_http.status_code = 500
        => trace = erroneous (114/0)
```

Figure 7: The outcome of Algorithm B for the dataset of Customer I.

The first rule refers to the span ”ctp.prov-repl.call-remedy” which was shown previously as the outcome of Algorithm A. Algorithm B reveals one of its important tags ”_annotations__spanLogs” with the corresponding value ”true.” The coverage of this rule is 95% (see Table 2).

The Rules of Figure 7	Coverage	Confidence
Rule #1	0.95	1
Rule #2	0.05	1

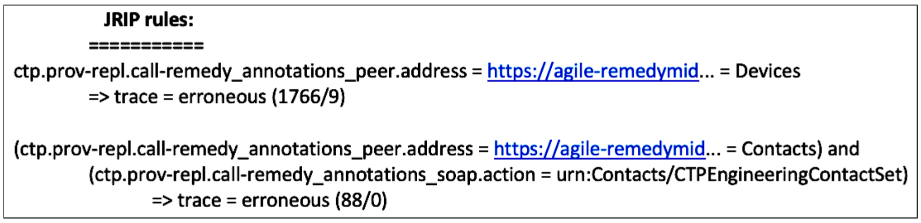
Table 2: The coverage and confidence of the rules of Figure 7.

The coverage of the second rule is around 5% (see Table 2). JRip also listed other rules with even smaller coverage. We are not showing them. The first rule has ideal coverage and confidence scores for sending to an end-user to explain the microservice performance degradation. The second rule should be excluded due to its small coverage.

It is possible that the revealed rules are not helpful in problem resolution. Although they are perfectly correlated with the errors, direct programming may have resulted. For example, the developers of applications are adding the tag ”error” with the value ”true” for a span if its ”status_code” is ”500” or the value of ”_spanLogs” is ”true.” That is why the confidence of those rules is very high.

We recommend the application of RIPPER in iterations by removing the spans and tags that appeared in the previous rule from the corresponding dataframes and reapplying the algorithm. We may continue until some stopping criteria are accomplished. One of the restrictions can be the time of executions. The second criterion can be the small values for coverage or confidence of the previous rules.

Returning to the previous example, we remove two features in the rules of Figure 7. The following Figure 8 shows the rules after the reiteration of the procedure. Information contained in those rules should be more helpful.



```
            JRIP rules:
        ===========
ctp.prov-repl.call-remedy_annotations_peer.address = https://agile-remedymid... = Devices
        => trace = erroneous (1766/9)

(ctp.prov-repl.call-remedy_annotations_peer.address = https://agile-remedymid... = Contacts) and
        (ctp.prov-repl.call-remedy_annotations_soap.action = urn:Contacts/CTPEngineeringContactSet)
              => trace = erroneous (88/0)
```

Figure 8: The outcome of Algorithm B for the dataset of Customer I after the second iteration.

They are showing some specific peer addresses somehow related to the errors. The first rule covers 73% and confidence 99.5% (see Table 3). The second rule has a small coverage but 100% confidence (see Table 3). We can send the first rule to a user to validate the insight.

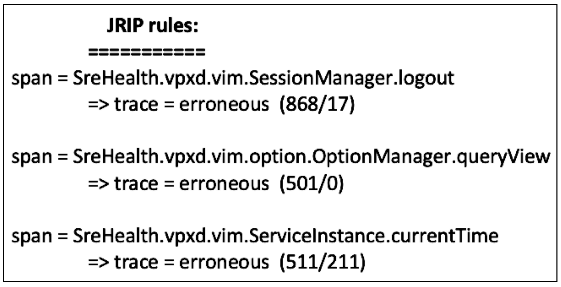
The Rules of Figure 8	Coverage	Confidence
Rule #1	0.73	0.995
Rule #2	0.04	1

Table 3: The coverage and confidence of the rules of Figure 8.

It is reasonable to start with Algorithm A. It will take several seconds to return the insights. Then, we can continue with Algorithms B and C while a user inspects the first set of rules. This will decrease the waiting time for a user as a sensitive constraint.

5.2 Customer II

In this case, we selected 11894 traces from the trace traffic passing through a malfunctioning microservice. The number of distinct trace types is 99, much bigger than in the previous example. Similarly, the number of distinct spans is also very big. We found 186 such spans. Algorithm A will work with a dataframe composed of 187 columns and 11894 rows. We performed data labeling via erroneous traces. We detected 2020 erroneous traces and 9873 normal ones. The execution time of Algorithm A is up to a second with JRip implementation. Figure 9 shows the first three rules.



```
          JRIP rules:
          ===========
span = SreHealth.vpxd.vim.SessionManager.logout
          => trace = erroneous  (868/17)

span = SreHealth.vpxd.vim.option.OptionManager.queryView
          => trace = erroneous  (501/0)

span = SreHealth.vpxd.vim.ServiceInstance.currentTime
          => trace = erroneous  (511/211)
```

Figure 9: The outcome of Algorithm A for the dataset of Customer II.

The first rule has 42% coverage and 98% confidence (see Table 4). It recommends the span "...SessionManager.logout". The second rule covers 25% and confidence 100% (see Table 4). It recommends "...OptionManager.queryView". The last rule has unacceptable 59% confidence (see Table 4). We will recommend only the first two rules for further consumption.

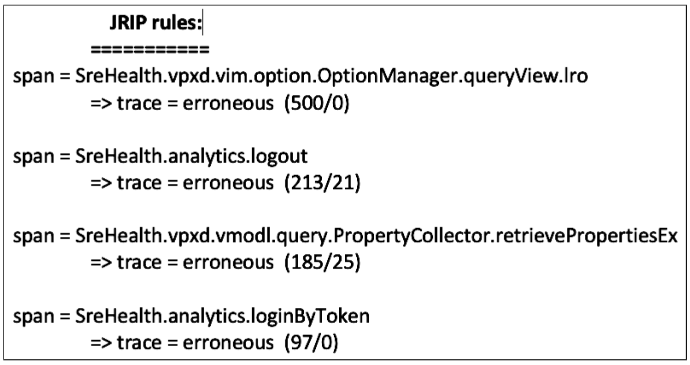
The Rules of Figure 9	Coverage	Confidence
Rule #1	0.42	0.98
Rule #2	0.25	1
Rule #3	0.15	0.59

Table 4: The coverage and confidence of the rules of Figure 9.

We continue with the second iteration after removing all features that appeared in the first iteration. Figure 10 reveals the new set of rules.

The first rule has ideal confidence and 25% coverage (see Table 5). We can add it to the previous list of recommendations. We can skip the remaining rules due to the small coverage (see Table 5).

We can continue with several iterations for each rule, calculating the scores and selecting the ones with sufficient coverage and confidence. Then, we can sort the rules by confidence in decreasing order and reveal the top five recommendations to a user. As mentioned above, the list of equivalent recommendations can be long. The best solution is to incorporate user feedback in the process of dataframe construction. Users can outline those spans or tags that can be helpful to see in the insights. It can help to decrease the number of columns of the dataframe and optimize the resource consumption by JRip.



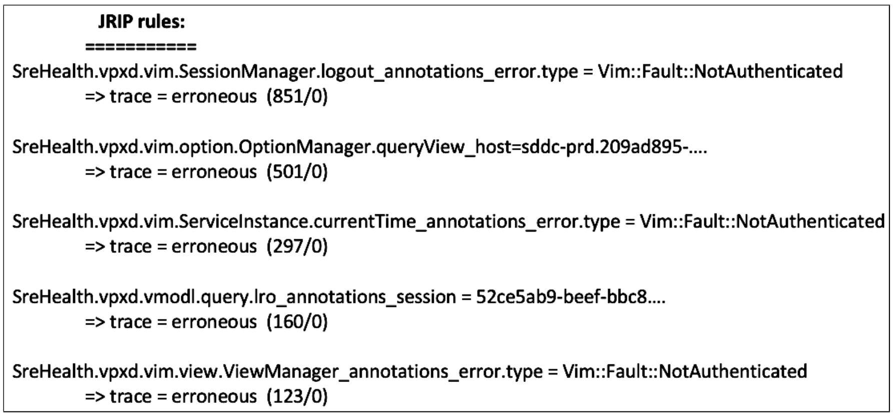
```
        JRIP rules:
       ===========
span = SreHealth.vpxd.vim.option.OptionManager.queryView.lro
       => trace = erroneous  (500/0)

span = SreHealth.analytics.logout
       => trace = erroneous  (213/21)

span = SreHealth.vpxd.vmodl.query.PropertyCollector.retrievePropertiesEx
       => trace = erroneous  (185/25)

span = SreHealth.analytics.loginByToken
       => trace = erroneous  (97/0)
```

Figure 10: The outcome of Algorithm A for the dataset of Customer II after the second iteration.

The Rules of Figure 10	Coverage	Confidence
Rule #1	0.25	1
Rule #2	0.1	0.9
Rule #3	0.08	0.86
Rule #4	0.05	1

Table 5: The coverage and confidence of the rules of Figure 10.

Now, let us return to Algorithm B. The number of distinct tags is enormous for the dataframe of Customer II. We found such 3002 distinct tags. The labels are the same as for Algorithm A. The execution time of JRip is around 20 seconds. Figure 11 shows the outcome of Algorithm B.



```
        JRIP rules:
       ===========
SreHealth.vpxd.vim.SessionManager.logout_annotations_error.type = Vim::Fault::NotAuthenticated
       => trace = erroneous  (851/0)

SreHealth.vpxd.vim.option.OptionManager.queryView_host=sddc-prd.209ad895-....
       => trace = erroneous  (501/0)

SreHealth.vpxd.vim.ServiceInstance.currentTime_annotations_error.type = Vim::Fault::NotAuthenticated
       => trace = erroneous  (297/0)

SreHealth.vpxd.vmodl.query.lro_annotations_session = 52ce5ab9-beef-bbc8....
       => trace = erroneous  (160/0)

SreHealth.vpxd.vim.view.ViewManager_annotations_error.type = Vim::Fault::NotAuthenticated
       => trace = erroneous  (123/0)
```

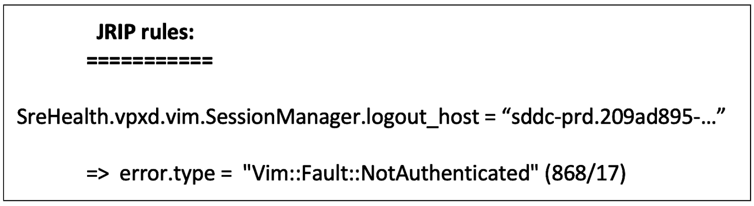
Figure 11: The outcome of Algorithm B for the dataset of Customer II.

We see interesting rules with 100% confidence (see Table 6). Three (the first, third, and fifth) directly indicate the same ”error.type” = ”NotAuthenticated”. The second indicates the specific host and the fourth refers to the session.

The Rules of Figure 11	Coverage	Confidence
Rule #1	0.42	1
Rule #2	0.25	1
Rule #3	0.15	1
Rule #4	0.08	1
Rule #5	0.06	1

Table 6: The coverage and confidence of the rules of Figure 11.

This example leads to another interesting data labeling idea. What if we want to explain the root cause of a specific error type? We can restart the labeling according to that specific value. For the example outlined in Figure 11, the label will take value 1 if ”error.type” = ”Vim::Fault::NotAuthenticated”, and value 0, otherwise. For this specific example, we found 851 erroneous traces corresponding to that ”error.type” and applied Algorithm C for details. Preliminarily, we removed all fields containing ”status.codes” and ”logouts” as the previous rules already indicated the problem description in those terms. Figure 12 explains the source of that error. It indicates that almost all errors are connected with a specific host ”sddc-prd.209ad...”.



```
JRIP rules:
===========

SreHealth.vpxd.vim.SessionManager.logout_host = “sddc-prd.209ad895-...”

=> error.type = "Vim::Fault::NotAuthenticated" (868/17)
```

Figure 12: The outcome of Algorithm C for the dataset of Customer II, where data labeling was performed via ”error.type” = ”Vim::Fault::NotAuthenticated” values.

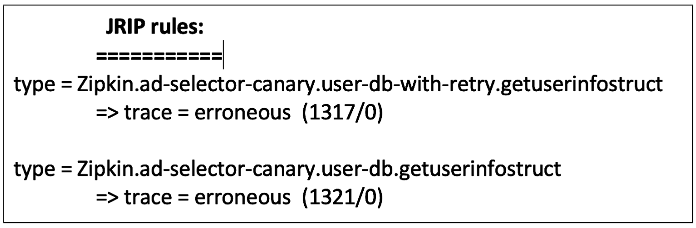
It is a valuable insight for a system administrator, leading to the direct source of the problem. A more general multi-class classification problem can be solved by putting different labels on all available values for a specified tag. The RCA will explain the origin of traces with the specific tag values via trace types, spans, or other tag values in a dataset related to a problem. Fortunately, RIPPER is dealing with multi-class problems with the same efficiency.

6 Customer III

We selected 5899 traces from the corresponding trace traffic and detected 26 distinct trace types. We also found 451 distinct spans and 4465 distinct tags. The number of distinct spans was rather huge. We verified the potential of Algorithm C for this scenario. The corresponding dataframe had 4917 columns and 5899 rows. We performed labeling based on the errors. We found 2754 erroneous and 3145 normal traces. The execution time of Algorithm C was 15 seconds. Actually, quite acceptable even for this big dataframe.

In general, the execution time depends on the number of distinct spans and tags. However, worth noting that it also depends on the number of classes in the categorical variables (the distinct values of tags). A bigger number of classes leads to longer execution times. Another important influential characteristic is noise. It is well known that RIPPER needs more time to construct rules in case of noisy datasets.

Figure 13 reveals that the most important component for explaining the errors of traces is the type (root span of a trace).



```
        JRIP rules:
        ===========
type = Zipkin.ad-selector-canary.user-db-with-retry.getuserinfostruct
        => trace = erroneous  (1317/0)

type = Zipkin.ad-selector-canary.user-db.getuserinfostruct
        => trace = erroneous  (1321/0)
```

Figure 13: The outcome of Algorithm C for the dataset of Customer III.

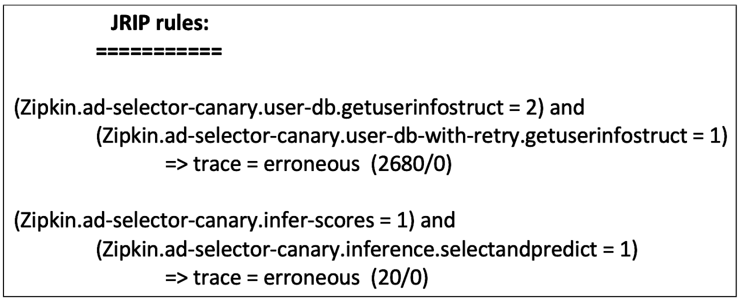
It shows two important types ”Zipkin.ad-selector-canary.user-db-retry.getuserinfostruct” and ”Zipkin.ad-selector-canary.user-db.getuserinfostruct”. Both processes are connected to a database. It means that the performance degradation of a microservice is connected with the DB. Table 7 presents the corresponding coverage and confidence of the rules.

The Rules of Figure 13	Coverage	Confidence
Rule #1	0.48	1
Rule #2	0.48	1

Table 7: The coverage and confidence of the rules of Figure 13.

There is an interesting modification of algorithms connected with spans. The idea is to modify the corresponding dataframes by considering the number of repetitions of the same spans in a trace. Until now, we have counted each span only once, no matter how many times a specific span appeared in the same trace. The entries of the dataframe were ”1” or ”NaN.” Now, the entries will be ”the number of appearances” or ”NaN.” This modification can tremendously strengthen the algorithms as erroneous micro-processes have a habit of repetitions.

For example, assume a problem with a credit card payment. In case of some malfunctioning, a user will repeat the payment, hoping to pay successfully, and the span corresponding to the payment will appear several times. This powerful modification is relatively simple to show for Customer III. Application of Algorithm A returns an empty set of rules, meaning explaining the erroneous traces via spans is impossible. However, the modification of Algorithm A reveals some insights shown in Figure 14.



```
          JRIP rules:
         ===========

(Zipkin.ad-selector-canary.user-db.getuserinfostruct = 2) and
         (Zipkin.ad-selector-canary.user-db-with-retry.getuserinfostruct = 1)
                  => trace = erroneous  (2680/0)

(Zipkin.ad-selector-canary.infer-scores = 1) and
         (Zipkin.ad-selector-canary.inference.selectandpredict = 1)
                  => trace = erroneous  (20/0)
```

Figure 14: Modification of Algorithm A by incorporation of span-repetitions in a trace for Customer III.

The first rule has a large coverage and 100% confidence (see Table 8). Interestingly, it is similar to the insights of the previous figure.

The Rules of Figure 14	Coverage	Confidence
Rule #1	0.97	1
Rule #2	0.007	1

Table 8: The coverage and confidence of the rules of Figure 14.

7 Trace-Latency Based RCA

Labeling traces based on an ”error” tag is a natural possibility, resulting in a powerful explanation engine. Another essential identifier of service performance degradation is the traces’ duration (latency). Usually, the execution time of a service has some average duration. Very short (usually with a zero duration) or long traces/spans indicate a shift from the normality and should be explained.

We outline two different approaches for the labeling of traces via durations. First, we need to separate traces into types (specific processes) and collect examples from different groups. If the number of examples in each group is rather significant, then p_{05} and p_{95} quantiles (or any other reasonable ones) of each group can serve as thresholds for the detection of small and large latencies. All traces in a group with durations smaller than p_{05} will get label -1, with durations longer than p_{95} will get label 1, otherwise label 0 (as normal traces). Then, RIPPER will explain separately both abnormal latency groups of traces. Those calculations can be made very efficient using the t-digest approach (see [Dunning and Ertl, 2019]), which can accurately estimate extreme quantiles via a streaming approach based on traces collected over several days or months.

If the number of samples in each group is small, then the suggested method will always detect some false positives. In that case, we need a simple outlier detection method. For example, we can apply the whiskers’ method, which defines thresholds by the formulae

$$upper = q_{75} + 1.5 * (q_{75} - q_{25}),$$

and

$$lower = q_{25} - 1.5 * (q_{75} - q_{25}).$$

Figure 15 shows how the whiskers' method can be used for an outlier detection for a specific trace type.

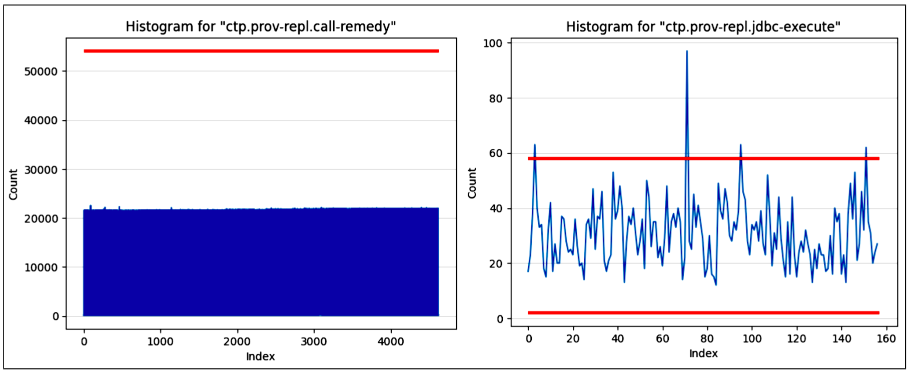
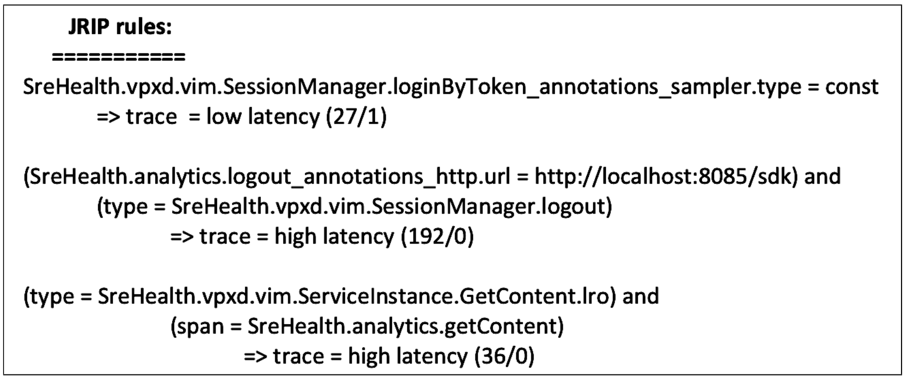


Figure 15: The whiskers' method for outlier detection in trace types.

Red lines correspond to the "upper" and "lower" thresholds calculated for each group. Each figure shows the distribution of trace durations (titles of images indicate the root spans). The left figure shows relatively compact durations of traces in that group, and the whiskers' method will not detect outliers. Some traces on the right figure violate the "upper" threshold.

Algorithms A, B, and C will work precisely similarly. The only difference is in labels; instead of errors, now we are explaining the durations of processes. Let us consider a dataset from the environment of Customer II. Labeling traces based on latency has resulted in 31 low-duration, 1127 high-duration, and 10735 normal traces. Figure 16 explains those outlying traces.



```
    JRIP rules:
   ===========
SreHealth.vpxd.vim.SessionManager.loginByToken_annotations_sampler.type = const
        => trace  = low latency (27/1)

(SreHealth.analytics.logout_annotations_http.url = http://localhost:8085/sdk) and
        (type = SreHealth.vpxd.vim.SessionManager.logout)
                => trace = high latency (192/0)

(type = SreHealth.vpxd.vim.ServiceInstance.GetContent.lro) and
                (span = SreHealth.analytics.getContent)
                        => trace = high latency (36/0)
```

Figure 16: Latency-based RCA for Customer II via Algorithm C.

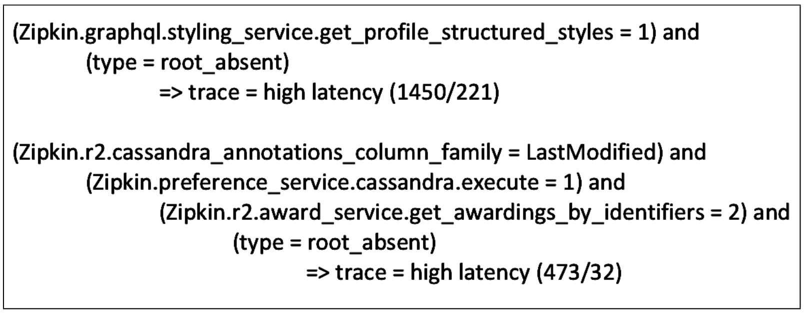
We removed some of the poor rules with small coverage and confidence. The first rule in Figure 16 explains the low-duration traces. It correctly explains 26 occurrences from 31 erroneous traces. The second two rules describe high-duration traces. They have low coverage but perfect confidence (see Table 9).

The Rules of Figure 16	Coverage	Confidence
Rule #1	0.84	0.96
Rule #2	0.17	1
Rule #3	0.03	1

Table 9: The coverage and confidence of the rules of Figure 16.

Now, let us consider a dataset from the environment of Customer III. As we mentioned before, we detected 8 different trace types. For each of them, we performed outlier detection and found 3027 erroneous traces with high latency (the label is ”1”), 16 erroneous traces with low latency (the label is ”-1”), and 6957 normal ones (the label is ”0”).

The rules of Figure 17 explain only the traces with high latency. RIPPER didn’t find patterns for the low-latency traces. We applied modified Algorithm A, which uses the number of appearances of spans in traces. Some traces in trace traffic have a root span. The root span defines the type of trace. However, some traces are arriving without the root span. We are collecting them in a unified group named ”root_absent.” We can remove previously appeared spans and tags and reiterate the procedure for more useful insights.



```
(Zipkin.graphql.styling_service.get_profile_structured_styles = 1) and
        (type = root_absent)
                => trace = high latency (1450/221)

(Zipkin.r2.cassandra_annotations_column_family = LastModified) and
        (Zipkin.preference_service.cassandra.execute = 1) and
                (Zipkin.r2.award_service.get_awardings_by_identifiers = 2) and
                        (type = root_absent)
                                => trace = high latency (473/32)
```

Figure 17: Latency-based RCA analysis for Customer III via modified Algorithm A.

Table 10 presents the coverage and confidence of the rule in Figure 17. The confidences of the rules are acceptable. The coverage of the first rule is also acceptable.

8 Filtering of RCA Recommendations via Baseline Estimation

Validation of recommendations before assigning them to an IT specialist for further remediation actions is an important milestone. We have set up several importance scores

The Rules of Figure 17	Coverage	Confidence
Rule #1	0.4	0.85
Rule #2	0.15	0.93

Table 10: The coverage and confidence of the rules of Figure 17.

for the prioritization of rules. However, there can be more general sources of noise that are worth further consideration. Many applications carry background noise (a bunch of erroneous traces), which does not impact the performance. The corresponding trace traffic contains a bucket of erroneous traces, and RIPPER can explain them with relevant rules. However, those rules are misleading for the final goal of recovering the performance degradation. More valuable insights will be recovered from those erroneous traces resulting from a change in trace traffic. Here, we suggest a procedure that can help to filter out the recommendations corresponding to an application background noise.

Change detection is possible while comparing two portions of trace traffic. The first one corresponds to a time window without any performance issues. We call it as peacetime period. Worth noting once more that RIPPER applied to this trace traffic may result in several strong rules that we want to eliminate from the final list of recommendations. The second one corresponds to a period with some IT issues. We call it as wartime period. We want the explanation of the erroneous traces in the wartime period subject to the peacetime one (conditional RCA). We call this approach war-peace-time RCA.

A naïve approach to war-peace-time RCA would assume the application of JRip separately to both dataframes with a filtering procedure for the wartime rules subject to the peacetime rules. For example, we can exclude those spans and tags from the wartime rules that simultaneously appear in the peacetime rules. However, we don't have a consistent procedure for this task. We suggest another automatic procedure that directly retrieves the required rules, excluding the process of manual filtering.

The first step is the application of RIPPER to the peacetime period. We need only the trained classification model without rules. The second step is applying the trained model to the wartime period. We are interested only in the misclassified traces we use for trace relabelling. All erroneous traces in the wartime period that were classified correctly will change their labels from "1" (default label for the erroneous traces) to "0". We are not interested in the correctly classified erroneous traces, as they probably already appeared in the peacetime dataset. We keep only those labels that the peacetime model misclassified. Finally, we apply RIPPER to the relabelled wartime dataframe and derive the required rules. This procedure can be used for three Algorithms: A, B, and C. The entire procedure can be reiterated as before.

Let us consider war-peace-time RCA for a dataset from the Customer I environment. Algorithm B applied to a peacetime period revealed the importance of the tag "http.status_code = 500" associated with the span "ctp.remedywebapp.POST/arsys/services-/ARService". We have already seen this recommendation before. The rule has 45.4% coverage and 100% confidence. Application of RIPPER to a wartime period revealed the same recommendation with 100% confidence and 49.6% coverage. However, the war-peace-time RCA returns an empty list of recommendations after filtering some weak rules. It means that war-peace-time RCA is capable of natural filtering of redundant rules. The main drawback of this approach is the difficulty of selecting the corresponding periods for analysis.

9 Conclusions and Future Work

Troubleshooting native cloud applications requires profound domain knowledge and essential skills due to the complexity of distributed systems with nonlinear probabilistic interrelations mostly hidden without modern APM solutions. The latest collects and stores all available information like time series, logs, traces, and events for detection, identification, and remediation of IT issues. Timely resolution of performance degradations is realistic only with ML/AI-empowered data analytics. However, system administrators require interpretable /explainable innovative solutions otherwise, no one will blindly follow black-box recommendations.

The paper's main goal is to nourish distributed tracing with explainable RCA for accelerating the resolution of performance degradations. Distributed tracing is a method of application monitoring based on microservices architecture. It provides better visibility of application components compared to classical monitoring based on time series and logs. The atoms of information are traces that describe the requests flowing through the different microservices of an application. They comprise spans and tags for more detailed monitoring of hidden processes. The analysis of a malfunctioning microservice can be performed via trace-traffic passing through that component.

We considered applying a classification rule-induction approach known as RIPPER to the trace traffic to reveal the explanations of performance degradations that can accelerate the mean time to resolution in such complex environments. The explanations can be derived from rules containing information regarding the spans, tags, and their values. We suggested three different algorithms that swept the trace traffic with different resolutions. Algorithm A worked with the lowest resolution, incorporating trace types and spans. Algorithm B utilized the tags and their values. Algorithm C combined both approaches. The latest internally selected the required level of resolution.

We suggested trace labeling based on different indicators. One of the approaches performed data labeling based on the internal tag known as ”error.” Its value differentiates typical and erroneous traces. Another approach utilized trace durations, assuming that similar trace types had almost identical durations, and violation of that principle should be the reason for the inspection. Finally, we considered data labeling based on specific tag values of spans that should be interesting to explain different tag categories, like ”error.type.”

We experimented with customer data and illustrated the benefits of different scenarios. We proved that the outcomes of RIPPER were highly explainable, and the execution times were acceptable, even for trace traffic with a large number of spans and tags. RIPPER has technical advantages like tolerance towards missing values, flexibility to work with multi-class problems, and the capability to consume numeric and categorical variables. This approach was a part of patented analytics designed for native cloud applications. It was productized by VMware and passed multi-layer validation steps.

The explainability and predictive power of ML algorithms have opposite directions. As a highly explainable approach, RIPPER has comparatively lower predictive power than neural networks, boosting, SVM, and others. In some situations, the classification accuracy will be insufficient for relying on the corresponding recommendations. We will investigate applying more powerful approaches to the problem, which can predict the appearance of erroneous traces with more extensive coverage and confidence. However, due to a lack of explainability, we must explore XAI possibilities (see [Harel et al., 2022, Hooker et al., 2018]) to enhance this disadvantage. Probable solutions can be the application of the SHAP method (see [Lundberg and Lee, 2017, Mayer, 2022, Shapley, 1953]), LIME (see [Ribeiro et al., 2016]), built-in feature importance analysis in the boosting methods (see [Sandri and Zuccolotto, 2008]) and model agnostic permutations-based approach (see [Breiman, 2001, Altmann et al., 2010]). Another interesting approach is incorporating user feedback into the process of recommendation prioritization.

Acknowledgements

The RA Science Committee funded Arnak Poghosyan in the frames of the research project № 20TTAT-AIa014. The research is conducted within the ADVANCE Research Grants provided by the Foundation for Armenian Science and Technology.

We thank anonymous reviewers for the critical reading of our paper. Their insightful comments helped to improve the manuscript.

References

[Agrawal et al., 1993] Agrawal, R., Imieliński, T., and Swami, A. (1993). Mining association rules between sets of items in large databases. In *Proceedings of the 1993 ACM SIGMOD international conference on Management of data*, pages 207–216.

[Altmann et al., 2010] Altmann, A., Toloşi, L., Sander, O., and Lengauer, T. (2010). Permutation importance: a corrected feature importance measure. *Bioinformatics*, 26(10):1340–1347.

[Barredo Arrieta et al., 2020] Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., and Herrera, F. (2020). Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, 58:82 – 115.

[Breiman, 2001] Breiman, L. (2001). Random forests. *Mach. Learn.*, 45(1):5–32.

[Cai et al., 2019] Cai, Z., Li, W., Zhu, W., Liu, L., and Yang, B. (2019). A real-time trace-level toot-cause diagnosis system in Alibaba datacenters. *IEEE Access*, 7:142692–142702.

[Cohen, 1995] Cohen, W. W. (1995). Fast effective rule induction. In *Proceedings of the Twelfth International Conference on Machine Learning*, pages 115–123. Morgan Kaufmann.

[Dunning and Ertl, 2019] Dunning, T. and Ertl, O. (2019). Computing extremely accurate quantiles using t-digests.

[Elsaadawy et al., 2019] Elsaadawy, M., Kemme, B., and Younis, M. (2019). Enabling efficient application monitoring in cloud data centers using sdn.

[Fürnkranz, 1997] Fürnkranz, J. (1997). Pruning algorithms for rule learning. *Mach. Learn.*, 27(2):139–172.

[Fürnkranz et al., 2012] Fürnkranz, J., Gamberger, D., and Lavrač, N. (2012). *Foundations of rule learning*. Cognitive Technologies. Springer, Heidelberg. With a foreword by Geoffrey I. Webb.

[Fürnkranz and Kliegr, 2015] Fürnkranz, J. and Kliegr, T. (2015). A brief overview of rule learning. In Bassiliades, N., Gottlob, G., Sadri, F., Paschke, A., and Roman, D., editors, *Rule technologies: Foundations, tools, and applications*, pages 54–69, Cham. Springer International Publishing.

[Fürnkranz and Widmer, 1994] Fürnkranz, J. and Widmer, G. (1994). Incremental reduced error pruning. In *Proc. 11th International Conference on Machine Learning*, pages 70–77. Morgan Kaufmann.

[Han et al., 2004] Han, J., Pei, J., and Yin, Y. (2004). Mining frequent patterns without candidate generation: A frequent-pattern tree approach. *Data Mining and Knowledge Discovery*, 8:53–87.

[Harel et al., 2022] Harel, N., Gilad-Bachrach, R., and Obolski, U. (2022). Inherent inconsistencies of feature importance.

[Harutyunyan et al., 2022] Harutyunyan, A., Aghajanyan, N., Harutyunyan, L., Poghosyan, A., Bunarjyan, T., and Han Vinck, A. (2022). On diagnosing cloud applications with explainable ai. In Hajian, A., Baloian, N., Inoue, T., and Luther, W., editors, *Third CODASSCA Workshop, Yerevan, Armenia: Collaborative Technologies and Data Science in Artificial Intelligence Applications*, pages 23–26, Berlin. Logos Verlag.

[Harutyunyan et al., 2020a] Harutyunyan, A. N., Grigoryan, N. M., and Poghosyan, A. V. (2020a). Fingerprinting data center problems with association rules. In Hajian, A., Baloian, N., Inoue, T., and Luther, W., editors, *Proceedings of the Second CODASSCA Workshop, Yerevan, Armenia: Collaborative Technologies and Data Science in Artificial Intelligence Applications*, pages 152–158, Berlin. Logos Verlag.

[Harutyunyan et al., 2020b] Harutyunyan, A. N., Grigoryan, N. M., Poghosyan, A. V., Dua, S., Antonyan, H., Aghajanyan, K., and Zhang, B. (2020b). Intelligent troubleshooting in data centers with mining evidence of performance problems. In Hajian, A., Baloian, N., Inoue, T., and Luther, W., editors, *Proceedings of the Second CODASSCA Workshop, Yerevan, Armenia: Collaborative Technologies and Data Science in Artificial Intelligence Applications*, pages 169–180, Berlin. Logos Verlag.

[Harutyunyan et al., 2019] Harutyunyan, A. N., Poghosyan, A. V., Grigoryan, N. M., Hovhannisyan, N. A., and Kushmerick, N. (2019). On machine learning approaches for automated log management. *J. Univers. Comput. Sci. (JUCS)*, 25(8):925–945.

[Harutyunyan et al., 2018] Harutyunyan, A. N., Poghosyan, A. V., Kushmerick, N., and Grigoryan, N. (2018). Learning baseline models of log sources. In Hajian, A., Luther, W., and Vinck, A. J. H., editors, *Proceedings of the CODASSCA Workshop, Yerevan, Armenia: Collaborative Technologies and Data Science in Artificial Intelligence Applications*, pages 145–156, Berlin. Logos Verlag.

[Heger et al., 2017] Heger, C., van Hoorn, A., Mann, M., and Okanović, D. (2017). Application performance management: State of the art and challenges for the future. In *Proceedings of the 8th ACM/SPEC on International Conference on Performance Engineering*, ICPE ’17, page 429–432, New York, NY, USA. Association for Computing Machinery.

[Helmer et al., 1998] Helmer, G., Wong, J., Honavar, V., and Miller, L. (1998). Intelligent agents for intrusion detection. In *Proceedings IEEE Information Technology Conference, Syracuse, NY*, pages 121–124. Springer.

[Helmer et al., 2002] Helmer, G., Wong, J. S., Honavar, V., and Miller, L. (2002). Automated discovery of concise predictive rules for intrusion detection. *Journal of Systems and Software*, 60(3):165–175.

[Hooker et al., 2018] Hooker, S., Erhan, D., jan Kindermans, P., and Kim, B. (2018). Evaluating feature importance estimates. *arXiv*.

[Hühn and Hüllermeier, 2009] Hühn, J. and Hüllermeier, E. (2009). FURIA: An algorithm for unordered fuzzy rule induction. *Data Min. Knowl. Discov.*, 19(3):293–319.

[John et al., 1994] John, G. H., Kohavi, R., and Pfleger, K. (1994). Irrelevant features and the subset selection problem. In *Machine Learning: Proceedings of the 11th International Conference*, pages 121–129. Morgan Kaufmann.

[Klaise et al., 2020] Klaise, J., Van Looveren, A., Cox, C., Vacanti, G., and Coca, A. (2020). Monitoring and explainability of models in production.

[Knoche and Hasselbring, 2019] Knoche, H. and Hasselbring, W. (2019). Drivers and barriers for microservice adoption – a survey among professionals in germany. *Enterprise Modelling and Information Systems Architectures (EMISAJ) – International Journal of Conceptual Modeling*, 14(1):1–35.

[Lee and Stolfo, 1998] Lee, W. and Stolfo, S. J. (1998). Data mining approaches for intrusion detection. In *Proceedings of the 7th Conference on USENIX Security Symposium - Volume 7*, SSYM'98, page 6, USA. USENIX Association.

[Lin et al., 2020] Lin, F., Muzumdar, K., Laptev, N. P., Curelea, M.-V., Lee, S., and Sankar, S. (2020). Fast dimensional analysis for root cause investigation in a large-scale service environment. *Proceedings of the ACM on Measurement and Analysis of Computing Systems*, 4(2):1–23.

[Lin et al., 2018] Lin, J., Chen, P., and Zheng, Z. (2018). Microscope: Pinpoint performance issues with causal graphs in micro-service environments. In *ICSOC*.

[Liu et al., 2021] Liu, D., He, C., Peng, X., Lin, F., Zhang, C., Gong, S., Li, Z., Ou, J., and Wu, Z. (2021). Microhecl: High-efficient root cause localization in large-scale microservice systems. In *2021 IEEE/ACM 43rd International Conference on Software Engineering: Software Engineering in Practice (ICSE-SEIP)*, pages 338–347.

[Liu and Motoda, 1998] Liu, H. and Motoda, H. (1998). *Perspectives of Feature Selection*, pages 17–41. Springer US, Boston, MA.

[Lundberg and Lee, 2017] Lundberg, S. and Lee, S.-I. (2017). A unified approach to interpreting model predictions.

[Mahmud et al., 2021] Mahmud, R., Ramamohanarao, K., and Buyya, R. (2021). Application management in fog computing environments. *ACM Computing Surveys*, 53(4):1–43.

[Mannila et al., 1995] Mannila, H., Toivonen, H., and Verkamo, A. I. (1995). Discovering frequent episodes in sequences extended abstract. In *Proceedings of the First International Conference on Knowledge Discovery and Data Mining*, KDD'95, page 210–215. AAAI Press.

[Marvasti et al., 2013] Marvasti, M. A., Poghosyan, A. V., Harutyunyan, A. N., and Grigoryan, N. M. (2013). Pattern detection in unstructured data: An experience for a virtualized IT infrastructure. In Turck, F. D., Diao, Y., Hong, C. S., Medhi, D., and Sadre, R., editors, *2013 IFIP/IEEE International Symposium on Integrated Network Management, IM 2013, Ghent, Belgium, May 27-31, 2013*, pages 1048–1053. IEEE.

[Mayer, 2022] Mayer, M. (2022). Shap for additively modeled features in a boosted trees model.

[Nag et al., 2021] Nag, D. A., Grigoryan, N. M., Poghosyan, A. V., and Harutyunyan, A. N. (2021). Methods and systems that identify dimensions related to anomalies in system components of distributed computer systems using traces, metrics, and component-associated attribute values. Patent US US11113174. Filed March 27, 2020. Issued Sep 7, 2021.

[Notaro et al., 2020] Notaro, P., Cardoso, J., and Gerndt, M. (2020). A systematic mapping study in aiops.

[Parker et al., 2020] Parker, A., Spoonhower, D., Mace, J., Sigelman, B., and Isaacs, R. (2020). *Distributed Tracing in Practice: Instrumenting, Analyzing, and Debugging Microservices*. O'Reilly Media, Incorporated.

[Poghosyan et al., 2021a] Poghosyan, A., Ashot, N., G.M., N., and Kushmerick, N. (2021a). Incident management for explainable and automated root cause analysis in cloud data centers. *JUCS - Journal of Universal Computer Science*, 27(11):1152–1173.

[Poghosyan et al., 2021b] Poghosyan, A., Harutyunyan, A., Grigoryan, N., Pang, C., Oganesyan, G., Ghazaryan, S., and Hovhannisyan, N. (2021b). An enterprise time series forecasting system for cloud applications using transfer learning. *Sensors*, 21(5).

[Poghosyan et al., 2016] Poghosyan, A. V., Harutyunyan, A. N., and Grigoryan, N. M. (2016). Managing cloud infrastructures by a multi-layer data analytics. In Kounev, S., Giese, H., and Liu, J., editors, *2016 IEEE International Conference on Autonomic Computing, ICAC 2016, Wuerzburg, Germany, July 17-22, 2016*, pages 351–356. IEEE Computer Society.

[Poghosyan et al., 2022] Poghosyan, A. V., Harutyunyan, A. N., Grigoryan, N. M., and Pang, C. (2022). Root cause analysis of application performance degradations via distributed tracing. In

Hajian, A., Baloian, N., Inoue, T., and Luther, W., editors, *Proceedings of the CODASSCA Workshop, Yerevan, Armenia: Collaborative Technologies and Data Science in Artificial Intelligence Applications*, pages 27–31, Berlin. Logos Verlag.

[Poghosyan et al., 2021c] Poghosyan, A. V., N., H. A., Grigoryan, N. M., Pang, C., Oganesyan, G., and Baghdasaryan, D. (2021c). Automated methods and systems that facilitate root cause analysis of distributed-application operational problems and failures. Patent US Application No.: 17/491,967 and 17/492,099. Filed Oct 1, 2021.

[Quinlan, 2014] Quinlan, J. R. (2014). *C4.5: programs for machine learning*. Elsevier.

[Ribeiro et al., 2016] Ribeiro, M. T., Singh, S., and Guestrin, C. (2016). ”why should i trust you?”: Explaining the predictions of any classifier.

[Sandri and Zuccolotto, 2008] Sandri, M. and Zuccolotto, P. (2008). A bias correction algorithm for the gini variable importance measure in classification trees. *Journal of Computational and Graphical Statistics*, 17(3):611–628.

[Shapley, 1953] Shapley, L. S. (1953). *17. A Value for n-Person Games*, pages 307–318. Princeton University Press, Princeton.

[Solé et al., 2017] Solé, M., Muntés-Mulero, V., Rana, A. I., and Estrada, G. (2017). Survey on models and techniques for root-cause analysis. ArXiv:1701.08546.

[Suriadi et al., 2013] Suriadi, S., Ouyang, C., van der Aalst, W., and ter Hofstede, A. (2013). Root cause analysis with enriched process logs. In Rosa, M. L. and Soffer, P., editors, *Business Process Management Workshops, International Workshop on Business Process Intelligence (BPI 2012). Volume 132 of Lecture Notes in Business Information Processing*, pages 174–186, Berlin. Springer-Verlag.

[Tzanettis et al., 2022] Tzanettis, I., Androna, C.-M., Zafeiropoulos, A., Fotopoulou, E., and Papavassiliou, S. (2022). Data fusion of observability signals for assisting orchestration of distributed applications. *Sensors*, 22(5).

[Vitali, 2022] Vitali, M. (2022). Towards greener applications: Enabling sustainable cloud native applications design.

[VMware, 2022] VMware (2022). Distributed tracing overview. https://docs.wavefront.com/tracing_basics.html. Accessed: 2022-12-1.

[Witten et al., 2005] Witten, I. H., Frank, E., Hall, M. A., and Pal, C. J. (2005). Practical machine learning tools and techniques. *Morgan Kaufmann*.

Journal of Universal Computer Science, vol. 29, no. 11 (2023), 1298-1318
submitted: 19/5/2023, accepted: 30/9/2023, appeared: 28/11/2023 CC BY-ND 4.0

Challenges and Experiences in Designing Interpretable KPI-diagnostics for Cloud Applications

Ashot Harutyunyan
(AI Lab at Yerevan State University, Yerevan, Armenia
Institute for Informatics and Automation Problems of NAS RA, Yerevan, Armenia
VMware, Palo Alto, US
https://orcid.org/0000-0003-2707-1039, aharutyunyan@vmware.com)

Arnak Poghosyan
(Institute of Mathematics of NAS RA, Yerevan, Armenia
VMware, Palo Alto, US
https://orcid.org/0000-0002-6037-4851, apoghosyan@vmware.com)

Lilit Harutyunyan
(VMware, Palo Alto, US
https://orcid.org/0000-0002-9558-9385, lharutyunyan@vmware.com)

Nelli Aghajanyan
(Deutsche Börse AG, Frankfurt am Main, Germany
https://orcid.org/0000-0002-3560-7502, nelli.aghajanyan@deutsche-boerse.com)

Tigran Bunarjyan
(Institute of Mathematics of NAS RA, Yerevan, Armenia
Technische Universität München, Munich, Germany
VMware, Palo Alto, US
https://orcid.org/0000-0003-4427-4284, tbunarjyan@vmware.com)

A.J. Han Vinck
(University of Duisburg-Essen, Duisburg, Germany
https://orcid.org/0000-0003-3437-3676, han.vinck@uni-due.de)

Abstract: Automated root cause analysis of performance problems in modern cloud computing infrastructures is of a high technology value in the self-driving context. Those systems are evolved into large scale and complex solutions which are core for running most of today's business applications. Hence, cloud management providers realize their mission through a "total" monitoring of data center flows thus enabling a full visibility into the cloud. Appropriate machine learning methods and software products rely on such observation data for real-time identification and remediation of potential sources of performance degradations in cloud operations to minimize their impacts. We describe the existing technology challenges and our experiences while working on designing problem root cause analysis mechanisms which are automatic, application agnostic, and, at the same time, interpretable for human operators to gain their trust. The paper focuses on diagnosis of cloud ecosystems through their Key Performance Indicators (KPI). Those indicators are utilized to build automatically labeled data sets and train explainable AI models for identifying conditions and processes "responsible" for misbehaviors. Our experiments on a large time series data set from a cloud application demonstrate that those approaches are effective in obtaining models that explain unacceptable KPI behaviors and localize sources of issues.

Keywords: Cloud infrastructures, Key Performance Indicators (KPI), automated root cause analysis (RCA), explainable ML/AI, feature importance, rule induction
Categories: I.2, K.6
DOI: 10.3897/jucs.112570

1 Intelligent Diagnosis of Cloud Environments

1.1 Explaining Factors of Misbehaviors

The general approach currently used by companies when it comes to RCA of performance issues in the customer ecosystems and own products is to rely on expertise of on-site users or support engineers. Expert efforts and knowledge are not anymore adequate for reliable management and quick remediation of misbehaving components of modern cloud environments tending to build self-driving capabilities (such as an application KPI optimization, the vision behind project Magna, https://blogs.vmware.com/virtualblocks/2019/08/26/vsan-project-magna/). Backtracking and finding root causes of failures in those distributed environments with high degree of sophisticated interrelations between data center objects is an unrealistic manual task for human operators.

Machine learning helps to automate the management of such complex systems [Josefsson 2017], [Sole et al 2017] that contain thousands of objects like VMs, Hosts, datastores, via monitoring millions of time series metrics, terabytes of logs, and application traces, to capture a high-resolution "image" of the entire stack. However, self-diagnostics of issues with those intelligent monitoring and analytics solutions (cloud vendors' products) is another fundamental problem at the customer environments requiring time-intensive analysis of support experts and substantial long-term investments. VMware Skyline [VMware Skyline 2022] summarizes common patterns (with the field experts involved) of product problems into remediation "rules" for more proactive support at the customer site later. A closed-loop global rule learning from products usage and performance data to maintain product KPIs healthy would be the advanced path several other companies currently follow (see HPE InfoSight [HPE 2022]).

Although data center management market progresses towards AI Ops solutions, like VMware Area (former vRelaize Operations Manager) [VMware Aria 2022], it still is providing only semi-automated root cause detection capabilities for customer applications under supervision, as well as for self-diagnosis, although enriched with various intelligent troubleshooting toolsets. In particular, they make finding evidence of potential causes of an alert or data center situation easier with discovery of "interesting" changes occurring in system events space, configuration properties, and data center flows for further user validation and decision making. Such a troubleshooting analytics [Harutyunyan et al 2020(1)] may apply statistical change point detection methods and entropic measures to derive the ranked lists of relevant patterns according to their importance. This kind of unsupervised approaches mitigate RCA problem but also produce false positive noise and redundancy. Full automation of RCA of an issue or its prediction remains unresolved. Our prior works ([Harutyunyan et al 2020(1)], [Poghosyan et al 2020(1)], [Harutyunyan et al 2020(2)], [Bunarjyan et al 2020], [Poghosyan et al 2020(2)], [Harutyunyan et al 2022], [Poghosyan et al 2022], [Baghdasaryan et al 2022]) reported in CODASSCA 2020 and 2022 workshops, extended papers ([Poghosyan et al 2021(1)], [Harutyunyan et al 2019], [Poghosyan et al 2021(2)]) in J.UCS and Sensors special issues on those workshops, earlier research ([Harutyunyan et al 2018], [Harutyunyan et al 2014]) on intelligent log analytics, represent various sorts of attempts at enabling and facilitating incident discovery and prediction, as well as RCA capabilities of cloud management solutions from different perspectives. In particular, the current paper builds on [Harutyunyan et al 2022].

Automated RCA with machine intelligence is a core problem in the self-driving data centers context [SDDC 2017]. However, for gaining user trust in ML solutions it is also essential and preferable to build such technologies on top of interpretable models [Barredo Arrieta et al 2020], [Fürnkranz et al 2012]. Core problems in reliable and intelligent cloud operations (including KPI diagnosis) are addressed in recent works ([Chen et al 2020], [Lyu et al 2021], [Lyu and Su 2023], [Wang et al 2023]) by various research groups.

There are multiple factors that hinder designing effective RCA solutions with ML for cloud computing infrastructures and applications, the main ones are

- lack or absence of labeled data;
- operator or expert verified/annotated/labeled data sets are hard to obtain in this domain and ungeneralizable from one environment to another because of ecosystem specifics.

Another aspect which is core to take into consideration is the explainability of an automated RCA. Industry is increasingly extending its frontiers with ML and AI, while facing the problem of explainability [Barredo Arrieta et al 2020] of sophisticated deep learning models and their outcomes. Although accurate ML models are valuable for automated RCA, their explainability is a desired feature to justify the reasons of failures and conditions that lead to such degradations. In addition to indicating actionable recommendations, those conditions are uncovering knowledge that can be leveraged in further optimization of the application. So, building trust between the user and AI should be an important requirement in designing data center diagnostics of the future. Therefore, our goal should be developing effective RCA methods which enable also explainable AI for products developed to manage cloud environments. That implies identifying ways to design intelligent systems that run on models with optimal trade-offs between their predictive power and explainability).

In that context, we outline ideas and a prototype solution for an automated RCA in terms of diagnosing KPI degradations that target troubleshooting customer data centers and/or cloud management products residing in their environments (thus enabling a proactive support while collecting high-frequency telemetry data from the products).

This paper presents some techniques and ML models that predict the potential causes of system's failures subject to its KPI, an underlying goal in the project Magna-Diagnosis mentioned above. Several regression and classification models were trained and analyzed using concepts of variable importance, decision trees and rule learners, as well as neural networks to identify and explain KPI degradations for a vRealize Operations deployment. These findings derived from high-accuracy ML models, which lack in human ground truth, were evaluated by the application developers to estimate their utility in practice and overall compliance with their expertise in long-term troubleshooting of the product issues.

Rule induction is always preferable if the interpretability of the models and patterns are required compared to their predictive power [Clark and Boswell 1998]. There are various rule induction algorithms [Fürnkranz et al 2012]. In this work we experimented with CN2 (see its implementation in the visual programming tool Orange [Orange 2023]).

In view of the above-mentioned challenges, our experience with designing relevant diagnostics pipeline proposal relies on the following building blocks:

1. Leveraging KPI metric as a source for generating labels for the entire data set of the application, while quantizing it into two or more class IDs;

2. Training regression and classification models to employ those in predicting KPI failures, while also evaluating relative variable/feature importance scores of those models to be utilized for explainability purposes;

3. Applying decision trees and rule induction (a form of explainable AI) algorithms to derive consistent conditions of KPI failures for a full KPI diagnosis and interpretability.

Methods are generic in nature and can be applied to different use cases, including proactive/predictive customer support for deployed cloud management instances or SaaS-based delivery of those services.

For an exemplary distributed application and its selected KPIs such as a latency metric, interpretable models are trained, validated against expertise of application developers, and used for producing run-time root cause recommendations on KPI abnormalities.

Overall, the objective of such a study is to identify important features and conditions of cloud applications subject to impact on KPIs. Based on this, intelligent cloud management solutions can provide recommender systems for optimizing applications performance (while indicating those important variables to be tuned) and predicting patterns causing unacceptable performance states (hence, accelerate the system recovery). This work focuses on experimental evaluation of the self-diagnostic use case of the technology leading cloud management solution VMware Aria Operations [VMware Aria 2022].

Figure 1 depicts this product application in its functions to monitor and guard multi-cloud infrastructures. The diagram reflects three cloud environments built on

1. VMware compute/storage/network virtualization solutions vSphere, vSAN, and NSX [VMware Prods 2023],

2. such an infrastructure hosted in Microsoft Azure, and

3. native Azure service, respectively.

While this distributed application is intelligently managing various cloud environments, the product itself might greatly benefit from self-healing capabilities or automated recommendations for performance improvements and recovery from misbehaviors. Architectural specifics and variety of workload patterns at different types of clouds may affect/stress vRealize Operations performance differently. Therefore, for self-diagnostics purposes, special models need to be trained for each case with specific requirements on KPIs behavior set by users.

Various ML algorithms are employed for comparative analysis including neural networks. Expert validation of discovered patterns promises wider adoptability of the approaches in real world scenarios with limited or unavailable annotated data sets.

1.2 Methodology Frameworks and Paper Sttructure

In our study we apply both regression and classification methods, including rule induction algorithms, as well as information-theoretic feature ranking techniques. In one scenario,

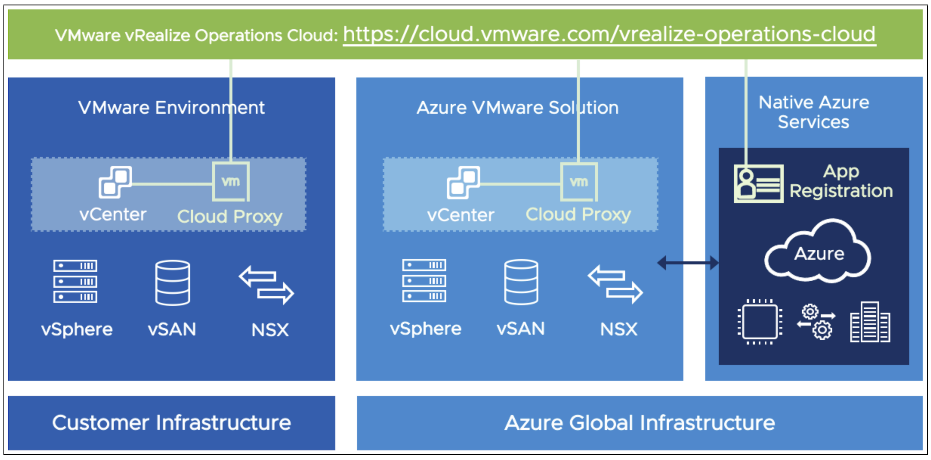


Figure 1: vRealize Operations in multi-cloud management.

we are interested in identifying potential factors explaining the KPI behavior, in another one, the problem is to predict KPI abnormality and deduce ”rules” leading to such situations.

Further research needs to address the problem of efficient management of multiple trade offing KPIs for an application.

The paper is organized as follows. Section 2 specifies use cases of application diagnostics from KPI perspectives and a sampled data set representing the performance monitoring of that application consisting of thousands of features. Section 3 focuses on experimental aspects of our investigations, while Section 4 expands on analysis of patterns. Section 5 shares validation results and challenges. Section 6 contains forward looking perspectives on this research and its productization with internal reviewer feedback on technology value.

2 Use Case of a Product KPI Diagnosis

We explain our methods on a data set measured by vRealize Operations (a multi-node distributed application consisting of several VMs) regarding its performance in terms of its self-monitoring metrics within a real deployment in one of our data centers. The product collects a large amount of such time series metrics that keep track of its own performance. They give a thorough picture of application’s state. The dataset we analyzed represents a collection of the application’s self-monitoring metrics for a node from a 6-node product deployment.

2.1 Dataset

The raw dataset was processed against gaps and normalized for further analysis. A node-specific data frame consisting of 3000 time series features (subject to 5 min regular monitoring interval) was considered, both numeric and categorical with 5100 instances each for a period of 18 days.

2.2 Key Performance Indicators

Different target variables (KPIs, see the relevant list [Self-Monitoring 2022] for vRealize Operations) in the dataset were used to generate labels for the rest of the data frame applying high quantile values of the metric. For training classification models, binary labels (normal vs abnormal state of the KPI) were generated using such an artificial and self-labeling technique. This way the dataset could be fed into classification algorithms for automated RCA to predict the positive/degraded class. The meaning behind this is to have ready ML models that could derive conditions/rules that interpret the degradation of KPI (positive class) or indicate the most influential dimensions/features for a long-term explainability of those degradations. To get better understanding of this approach, we have discretized the target variable to be 0 or 1 based on the 97, 95, 93, 90-quantiles of the KPI and picked up the quantile which resulted in the best performance.

New global KPIs have also been constructed since there are cases when a particular KPI cannot describe the required performance aspect accurately when taken separately. As node's behavior must be evaluated relative to the remaining nodes, taking a single KPI that reflects the relationship between only one to another node would not be descriptive. For each node from the 6-node product deployment, there are 5 self-monitoring metrics for each remaining node, which show the maximum/average response latency from the current node to others.

The super-metrics

Node|PingLatency|Max of Max

and

Node|PingLatency|Avg of Avg (Figure 2)

used for RCA are constructed by taking the maximum/average value among all 5 metrics' observations at a given point in time.

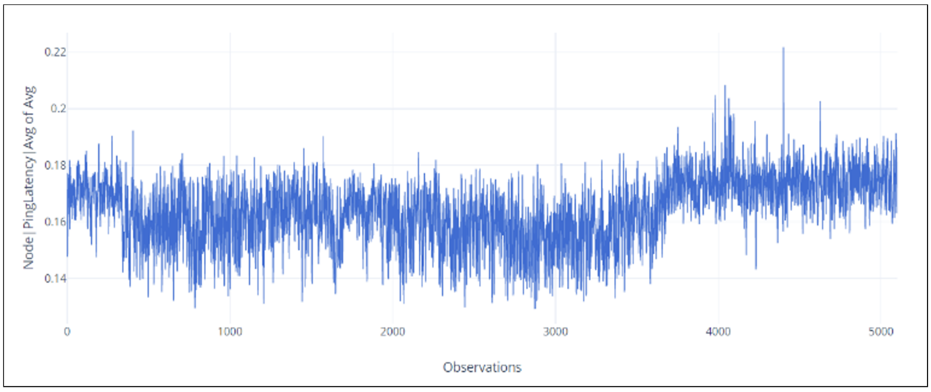


Figure 2: Plot for Node|PingLatency|Avg of Avg.

3 Experiments and Discussions

As mentioned above, we are interested in explaining a specific KPI degradation instance with classification and rule induction approaches, as well as in identifying flows highly influential on the KPI from historic perspectives (which implies employing regression settings). Both are important tasks. The first one explains an application situation, the second one derives factors that can be taken into account for application optimization planning. Moreover, highly important features and potential anomalies/outliers on them in a specific situation can also indicate reasons for a KPI failure.

In the regression analysis, a target variable (KPI) was chosen to be

Node|PingLatency|Max of Max.

k Nearest Neighbor (kNN) regression (to model expected non-linearity in the dataset) with parameter equal to 115 with cross-validation from the range $k = [5; 123]$ has yielded Root Mean Square Error (RMSE) equal to 0.119.

Another KPI was the constructed general KPI

Node|Ping Latency|Avg of Avg

shown in Figure 2. The kNN parameter was 15 with cross-validation from the range $k = [5; 123]$ and the algorithm has output RMSE equal to 0.062.

The third chosen KPI was

OverallThresholdChecking|MaxDuration,

which records the maximum duration for actions of those items that are used to process incoming observation data (against baseline thresholds for time series metrics set by the user or learned statistically). Here the number of neighbors was chosen to be 63 with cross-validation from the range $k = [5; 123]$ with RMSE equal to 0.027.

Table 1 summarizes top features subject to their relative importance values ("Coef" columns in the table) obtained for the kNN method and the corresponding KPIs (with short descriptors of those metrics as *Max of Max*, *Avg of Avg*, and *MaxDuration*, respectively). In terms of RMSE, the models performed well.

To benefit from supervised learning methods, we self-labeled the data frame with outlying behaviors of the KPI as positive class, using for that $97, 95, 93, 90$−quantiles of the metric in distinct experimental scenarios.

All above mentioned KPIs were considered to train different models (with several traditional classification algorithms) and compare their accuracies. However, the evaluation results for all those KPIs and ML algorithms were not satisfactory. The models did not have enough discrimination capacity to distinguish between positive and negative classes. This was a consequence of noisy dataset and class imbalance for the target variable. 90-quantile cut-off produced better results, so we considered this quantile for further improvement of the models.

We present a possible resolution to the problem of noise in the dataset below and the results obtained after improvements.

3.1 Neural Networks

Compared to traditional classification algorithms, Multi-Layer Perceptron (MLP) demonstrates a completely different predictive power on our data set. In particular, with 97-quantile cut-off for *Node|PingLatency|Avg of Avg*, MLP produces 96%-accuracy with

Precision=95% and Recall=0.96% in the binary classification between normal and abnormal states of the KPI. However, because of the interpretability issue of neural networks, we have to proceed with finding ways to improve our initial results on explainable models compromising the predictive power.

Max of Max		**Avg of Avg**		**MaxDuration**	
Feature	Coef	Feature	Coef	Feature	Coef
Node|CPU|Usage	100	*OverallThreshold Check|AvgDuration*	100	*OverallThreshold Check|AvgDuration*	100
Node|CPU1|User	97.53	*CassandraDB| LocalReadCount*	99.02	*Task|GetCommands |ElapsedTimeSum*	99.06
CPU|Usage	97.46	*Disk| FileSystem|writes*	98.88	*DAO|GetResource Metadata| AvgDuration*	98
Node|CPU2|User	97.28	*Network|Transmit Bytes*	98.34	*OverallThreshold Checking|Check Health|AvgDuration*	96
CPU17|User	95.81	*Mem|ActualFree*	98.30	*FeatureRequest| MaxDuration*	96.11
Node|CPU1| Combined	95.28	*Node|Memory| ActualUsed*	98.30	*Task|GetCommands |MaxElapsedTime*	93.58
Node|CPU2|Idle	95.26	*Node|Disk|DBReads*	97.97	*Call|Update ResourceRegion| AvgDuration*	82.21

Table 1: Important Variables for KPIs: Node|PingLatency|Max of Max, Node|PingLatency|Avg of Avg, and OverallThresholdChecking|MaxDuration.

3.2 Reduction of Noise

Class imbalance was a serious problem in the dataset. The considered target KPIs had only 510 positively labeled observations out of 5100 total number of observations. In attempts to achieve better performance, we experimented with several techniques such as undersampling the negative class by different ratios and performing feature selection by various approaches.

Undersampling techniques (removing of the instances in the majority class) has been applied by experimenting with two cases: 50%-positive-vs-50%-negative labels and 35%-positive-vs-65%-negative labels. AUC, Precision and Recall are improving in both cases for all KPIs and algorithms.

Although the results got better from undersampling, they still did not meet our expectations for reasonable analysis using classification models. To improve the results, we have applied several feature ranking/selection methods such as

- Gini Index (it quantifies how often a randomly chosen feature would be incorrectly classified if it were randomly assigned to a data point based on the distribution of the target variable);

- Information Gain (an entropy-based measure evaluating the ability of a feature to discriminate between different classes);
- ReliefF [Konotenko 1994] (it assesses the importance of features based on their ability to distinguish between neighboring instances with the same or different class labels);
- FCBF (Fast Correlation-based Filter) [Yu and Liu 2003]. This technique applies the idea of "predominant correlation". It gives a classifier-independent feature scoring mechanism by selecting features with high correlation with the target variable, but low correlation with other variables. For the correlation, it uses "symmetrical uncertainty" based on information theory and the concepts of entropy and information gain.

After numerous experiments, the results on FCBF ranked dataset outperformed the other models. FCBF ranking was applied after undersampling the negative class in the preceding experiments. When regression task is applied on the datasets with only FCBF non-zero best ranked features, RMSE metric does not substantially decrease, which indicates that both models have high accuracy.

3.3 Results with Undersampling and Feature Ranking

Several classification algorithms such as Decision Tree, CN2 rule induction, Logistic regression, and Naive Bayes were applied to our data set. Previous experiments showed that applying only one method of noise reduction is not enough on this dataset. Thereupon, both undersampling and FCBF ranking (which results in only from 24 to 30 features with non-zero score) were employed to get more predictive models with accurate findings. The dataset with 35% and 65% split (see Table 2 and 3 for two KPIs) performed substantially better, than 50%-vs-50% in terms of evaluation metrics for *OverallThresholdChecking|MaxDuration*, so this dataset will be used for further analysis.

Model	AUC	Precision	Recall
Tree	0.808	0.808	0.806
Naive Bayes	0.838	0.801	0.885
Logistic Regression	0.880	0.806	0.869
CN2 rule inducer	0.826	0.783	0.622

Table 2: Results for 90*-quantile positive labeling for OverallThresholdChecking|MaxDuration. Undersampled* 35%-*vs*-65%*, FCBF ranked.*

However, for *Node|PingLatency|Avg of Avg* there was no noticeable difference between the evaluation metrics for the two datasets, therefore the 35%-vs-65% ratio dataset was used for the experiment (Table 3). As we notice, 35%-vs-65% ratio for positive and negative class labels, respectively, renders the best results. So, further expansion of the positive class does not result in better performance of the models.

Model	**AUC**	**Precision**	**Recall**
Tree	0.684	0.667	0.635
Naive Bayes	0.834	0.765	0.767
Logistic Regression	0.696	0.602	0.759
CN2 rule inducer	0.769	0.728	0.652

Table 3: Results for 90*-quantile positive labeling for node|PingLatency|Avg of Avg. Undersampled* 35%-*vs*-65%*, FCBF ranked.*

3.4 Principal Component Analysis

Large datasets are often difficult to interpret. Principal component analysis (PCA) is a technique for reducing the dimensionality of these datasets, increasing interpretability, and simultaneously minimizing information loss [Jolliffe and Cadima 2016]. We have taken the KPI *OverallThresholdChecking|MaxDuration* with 90-quantile, ran PCA with number of components equal to 380 and got 95% explained variance. So, with only 380 features we can effectively find an optimal representation of the initial data set consisting of 3000 metrics. When comparing results on raw dataset with that of post-PCA, there are noticeable improvements observed. Undersampling the dataset obtained after applying PCA with 35% and 65% ratio and comparing results with the original dataset presented in Table 3, we observed that even the orthogonal components do not appropriately handle the noise. Therefore, there was no need to continue investigations with PCA on the initial data set.

4 Evaluation of Trained Models

To rigorously investigate the issues at KPIs

OverallThresholdChecking|MaxDuration

and

Node|PingLatency|Avg of Avg,

the concepts of variable importance, decision trees, and rule induction are applied. The derived candidate root cause features are compared among the models, intersections are observed, and a list of possible root cause metrics are presented for each KPI. In case of rule induction, when the conditions of the rule are met, i.e., when the features are constrained by the given values, the KPI degrades. So, for a KPI failure instance, the implementation needs to check which of rules are currently satisfied to recommend those for taking actions on.

4.1 Explaining Threshold Checking Duration

The rules corresponding to the positive class with highest Laplace quality obtained by CN2 rule induction are listed in Table 4.

Some rules have common features emphasizing the importance of the variable and its impact on the KPI. The distribution shows the number of observations that comply with the rules with target value equal to 1.

Rule	Quality	Distr

IF $ResourceSymptomRegionUpdate\|AvgDuration \geq 3.234$ ms
AND $SystemAttributes\|Health \geq 91\%$
THEN $OverallThresholdChecking\|\ MaxDuration = 1$
	0.938	14

IF $GetUpshards\|MaxDuration \geq 27$ ms **AND** $ResourceSymptomRegionUpdate\|AvgDuration \geq 1.4$ ms **AND** $SystemAttributes\|Health \leq 78\%$ **AND** $CassandraDB\|UsedLiveDiskspace \leq 892119$ *bytes* **AND** $SystemAttributes\|Health \geq 69\%$ **THEN** $OverallThresholdChecking\|MaxDuration = 1$	0.933	13

IF $DAO\|GetResourceMetadata\|AvgDuration \geq 1$ ms
AND $GetUpshards\|MaxDuration \geq 15$ ms
THEN $OverallThresholdChecking\|MaxDuration = 1$
	0.933	13

IF $ResourceSymptomRegionUpdate\|AvgDuration \geq 1.7$ ms **AND** $CapacityReclamationSettings\|MaxDuration \leq 9$ ms **AND** $CapacityReclamationSettings\|MaxDuration \leq 16$ ms **AND** $ControllerDB\|CPUSystem \geq 0.884$ KB **THEN** $OverallThresholdChecking\|MaxDuration = 1$	0.929	12
IF $CassandraDB\|UsedLiveDiskspace \geq 984121$ *bytes* **AND** $ResourceSymptomRegionUpdate\|AvgDuration \leq 0.776$ ms **AND** $SystemAttributes\|Health \geq 75\%$ **AND** $ResourceSymptomRegionUpdate\|AvgDuration \geq 0.721$ ms **THEN** $OverallThresholdChecking\|MaxDuration = 1$	0.917	10

IF $CassandraDB\|UsedLiveDiskspace \geq 984121$ *bytes*
AND $CapacityReclamationSettings\|MaxDuration \geq 23$ ms
THEN $OverallThresholdChecking\|MaxDuration = 1$
	0.875	6

Table 4: Results for OverallThresholdChecking|MaxDuration.

Overall, the obtained rules and the metrics participating have logical interpretation. The observations show that even in case when the system is healthy, but the average duration of resource symptom region update (a product-specific flow related to evaluating symptoms or unhealthy microstates/fragments at objects under monitoring) has high rate, the abnormality is inevitable. Another interesting rule stresses the importance of the

$$CassandraDB|UsedLiveDiskspace$$

threshold even when the duration is tolerable.

The decision tree for this KPI is presented in Figure 3.

If the metric

$$GetUpshards|MaxDuration$$

is greater than 14ms, then 135 out of 168 positively labeled KPI observations can be identified. The features that show duration are also beneficial for root cause analysis, as they reference exactly where the noticeable amount of time was wasted which resulted in the degradation of the KPI.

Another pattern, the metric

ResourceSymptomRegionUpdate|AvgDuraton

found on a branch of the decision tree may indicate that considerable amount of time was wasted on resource's symptom update, which may indicate that the number of symptoms is higher than usual.

The important variables found by Logistic regression and Naive Bayes for this KPI are presented in Table 5.

Besides last two features, the rest were present both in decision tree and CN2 rules, which implies that all four models have outputted the same important variables.

As of kNN regression, the list of important variables obtained from running the model on the raw data has only two matching metrics with this list. When running kNN regression algorithm on the dataset consisting of best ranked features, the number of matching metrics with high coefficients increases.

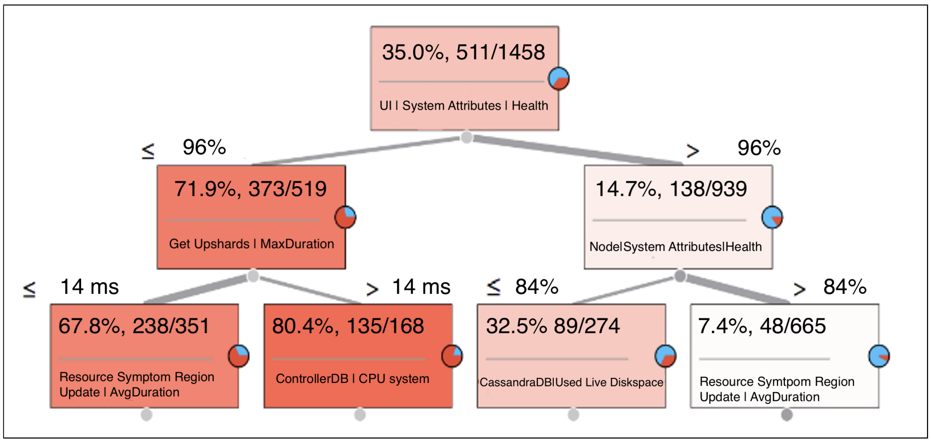

35.0%, 511/1458
UI I System Attributes I Health
≤ 96%
> 96%
71.9%, 373/519
Get Upshards I MaxDuration
14.7%, 138/939
NodeI System AttributesIHealth
≤ 14 ms
> 14 ms
≤ 84%
> 84%
67.8%, 238/351
Resource Symptom Region Update I AvgDuration
80.4%, 135/168
ControllerDB I CPU system
32.5% 89/274
CassandraDBIUsed Live Diskspace
7.4%, 48/665
Resource Symtpom Region Update I AvgDuration

Figure 3: Decision Tree for OverallThresholdChecking|MaxDuration.

CassandraDB|LiveDiskSpaceUsed
Get Upshards|MaxDuration
ResourceSymptomRegionUpdate|AvgDuration
ControllerSQL | CPUSystem
CapacityReclamationSettings | MaxDuration
GemfireClientCalls|LicensingService | ResponsesCount
ReclaimableVmsInfo|Count

Table 5: Important Variable for OverallThresholdChecking|MaxDuration by Logistic regression and Naive Bayes.

4.2 Explaining Node Latency

The rules corresponding to the positive class with highest Laplace quality obtained by CN2 rule induction are listed in Table 6.

It is interesting to notice that the first two rules have more than 30 examples supporting them. Besides high Laplace quality, this is a good measure to validate the accuracy of the results.

From the table it is clear that besides the features showing duration and elapsed time, other system metrics also have their effect on the KPI's behavior.

Metrics such as

$$Network|TransmitBytes$$

and

$$APIService|CurrentHeapSize$$

influence the KPI, as the former is the number of transmit bytes over the network, which naturally affects the latency between the nodes and the latter is the current heap size for API calls. This, interestingly enough, can impact on the latency between the nodes.

Rule	Quality	Distr
IF *Task|GetTokens|MinimumElapsedTime* $\geq$ 5 ms **AND** *Call|GetSetting|AvgDuration* $\geq$ 3.2 ms **AND** *Service|Total number of open file descriptors* $\geq$ 463 **AND** *Network|TransmitBytes* $\geq$ 61043428 **AND** *API|CurrentHeapSize* $\geq$ 201MB **THEN** *Node|PingLatency|Avg of Avg* = 1	0.971	32
IF *API|CurrentHeapSize* $\geq$ 409MB **AND** *OverallThresholdChecking|MaxDuration* $\geq$ 855 ms **AND** *Network|TransmitBytes* $\leq$ 115890416 **AND** *NewResourcesCount* $\geq$ 31 **AND** *Network|TransmitBytes* $\geq$ 71863520 **THEN** *Node|PingLatency|Avg of Avg* = 1	0.947	35

IF *CassandraDB|LocalReadCountDelta* $\geq$ 1803
THEN *Node|PingLatency|Avg of Avg* = 1
	0.917	10

IF *APICall|GetResourceRelationship|MinResponseTime* $\geq$ 73 ms
AND *APICall|GetResourceRelationship|MinResponseTime* $\leq$ 84 ms
AND *API|CurrentHeapSize* $\geq$ 378MB
THEN *Node|PingLatency|Avg of Avg* = 1
	0.938	14

IF *OverallThresholdChecking|MaxDuration* $\geq$ 763 ms
AND *Call|GetSetting|AvgDuration* $\geq$ 13.8 ms
THEN *Node|PingLatency|Avg of Avg* = 1
	0.929	12

Table 6: Rules for Node|pingLatency|Avg of Avg.

The decision tree for this KPI is presented in Figure 4.

The previous KPI

$$OverallThresholdChecking|MaxDuration$$

has a significant effect on the current KPI. If the max duration is greater than 755 milliseconds, with 75.4% probability the KPI is anomalous. Moreover, if the count of new resources (objects in the infrastructure) is higher than the number indicated in the tree, 393 out of 503 observations of KPI with positive label are identified. In the case, when the previous KPI is less than or equal to 755 milliseconds, the probability of abnormal behavior of this KPI is relatively small. Metrics like

$$CassandraDB|LocalReadCountDelta,$$
$$Network|TransmitBytes$$

have logical impact on the KPI whose abnormal behavior is directly affected by them. If the read count is large enough and the transmitted bytes between network exceed the threshold then our KPI has surely abnormal rate.

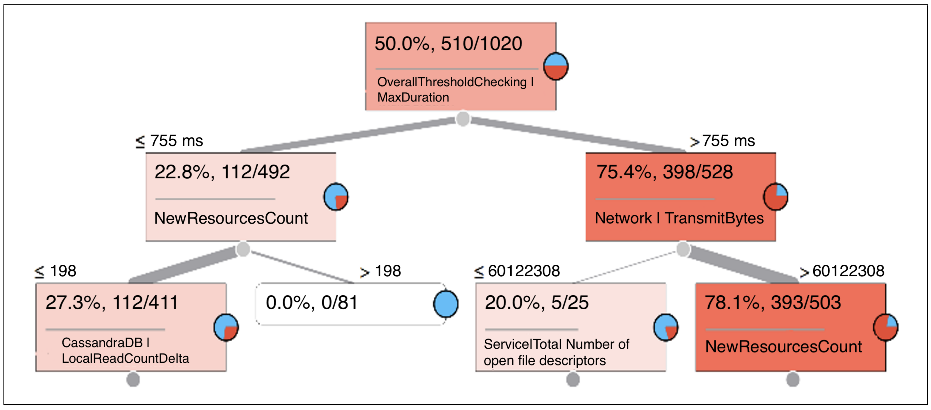


Figure 4: Decision Tree for Node|PingLatency|Avg of Avg.

The important variables found by the Logistic regression and Naive Bayes models for

$$Node|PingLatency|Avg\ of\ Avg$$

are presented in Table 7.

5 Initial Validation of Results

We have approached the problem of RCA from two perspectives. First, by using regression models we have analyzed the data to get long-term/history-based possible root cause metrics. The space is being narrowed down to some important variables, which impact the KPI's abnormal behavior the most. The kNN regression model applied both on raw and ranked data show that these metrics affect the

$$OverallThresholdChecking|MaxDuration$$

degradation the most:

ResourceSymptomRegionupdate|AvgDuration,
CassandraDB|UsedLiveDiskspace,
ControllerDB|CPUsystem,
GemfireClientCalls|LicensingService|RequestsCount.

When applying classification models such as rule induction and decision tree algorithms, incident-based approach is considered. The rules which violation causes KPI degradation are being discovered. So, in case of KPI misbehavior, the implementation can analyze the extracted rules and the metrics that compose those conditions, recognize if the conditions are met and localize the causes with regards to few metrics. The list of potential sources obtained by applying classification models are presented in Table 8. Getting metrics indicating durations of specific actions may seem straightforward, however, this shows exactly what actions consumed most of the time. In addition, getting specific bounds on these metrics show the thresholds which when violated will cause KPI misbehavior. Moreover, it turns out that the proportion of database's used disk space and CPU can negatively affect the amount of time consumed by the system to analyze new observations. All the results are logical and relevant, which are good measures for evaluating root causes.

CassandraDB | LocalReadCountDelta
OverallThresholdChecking | MaxDuration
Network | TransmitBytes
NewResourcesCount
API Service | CurrentHeapSize
Task | GetTaskStatuses | ResponsesRecieved
Task | GetTokens | MinimumElapsedTime
Task | ResourceRegistration | MaxElapsedTime
Call | GetSetting | AvgDuration
Service | Total number of open file descriptors

Table 7: Important Variables for Node|pingLatency|Avg of Avg by Logistic regression and Naive Bayes.

The kNN regression model applied both on raw and ranked data output the following list of important variables affecting the

Node|PingLatency|Avg of Avg

the most:

OverallThresholdChecking|AvgDuration,
Network|TransmitBytes,
FeatureRequest|MaxDuration,
API|CurrentHeapSize.

The list of potential root cause metrics for this KPI obtained from classification models are shown in Table 8.

For this KPI as well, the specific duration metrics are beneficial for troubleshooting the misbehavior of the KPI. An interesting result is the metric

$$NewResourcesCount$$

which shows that the count of new resources discovered by the application causes latency increase among the nodes. When particular thresholds for

$$APIService|CurrentHeapSize$$

and

$$CassandraDB|LocalReadCountDelta$$

are exceeded the misbehavior of the KPI is unavoidable.

The lists mentioned above are the potential sources for the KPIs degradation need to be checked first when that event occurs.

OverallThresholdChecking|MaxDuration	**Node|PingLatency|Avg of Avg**
CapacityReclamationSettings|MaxDuration	*CassandraDB|LocalReadCountDelta*
CassandraDB|LiveDiskSpaceUsed	*OverallThresholdChecking|MaxDuration*
Get Upshards|MaxDuration	*Network|TransmitBytes*
ResourceSymptomRegionUpdate|AvgDuration	*NewResourcesCount*
ControllerDB|CPUSystem	*Task|GetTokens|Minimum ElapsedTime*
	Call|GetSetting|AvgDuration
	Service|Total number of open file descriptors

Table 8: Metric summary for two KPIs (classification scenario).

5.1 Insights Learned

Based on thorough analysis of experimental observations, we get initial insights on the utility of the proposed approach:

- The techniques described in this paper can be used for finding root causes of any KPI degradation if the labeling of data is available or performed in a self-supervised way appropriately.
- The findings discussed above demonstrate that with a relevant dataset of self-metrics from the product deployments at customers, enough powerful models can be trained to predict and explain application/product misbehaviors.

– Tracking the complex conditions constituting the rules we can proactively measure the risk of application failures at the customer's side or in a SaaS delivery of those services. It means that the risk score of the KPI deterioration will increase along with the occurrence of conditions in the rules.

– The methods while productized into cloud management solutions (including their self-healing intelligence engines) can enhance application-aware depth and autonomous capabilities of those data-agnostic services with interpretable recommendations for human administrators.

5.2 Feedback from Product Experts

For rigorous validation of obtained models and rules, as well as importance features, extended studies/surveys on performance troubleshooting of the considered application for a long-term period are required. Such a study should rely on multiple models trained for specific KPI at various environments and quantifying relevant quality measures of recommendations, such as indicative relevance of rules, real importance of derived features in problem resolution, in general, a mean-time-to-repair rate (MTTR). However, setting up this test bed with its comprehensive evaluation over time remains another challenge which might be overcome with a pilot productization of the methods for a set of customers.

In our initial validation of the methodology, we adopted a different approach. The current findings were presented to a small group of experts experienced in troubleshooting this cloud application. The feedback was positive as most of the metrics found were also observed by experts as potential root causes of the given KPI's degradation in customer environments. Some of the metrics found by our analysis were not considered as root causes of the given KPI's anomalous behavior before, however, according to experts there is a logical connection between them, and those metrics should be further evaluated in multiple environments. The extracted rules were of interest to product experts, as they interpreted some of them and confirmed the correctness. Then the goal was to get from the evaluators an overall usefulness rate of discovered patterns. Such a score might attribute a high confidence to our analysis, as the jury consisted of the most experienced product developers and support engineers. This survey demonstrated an approximately 90% utility degree of recommendations presented to them (we need to take into account also that the approximation of real performance issues with generating labels from outlying KPI behaviors might be only close representation of those problems (ground truth)), while 10% remaining patterns were noted with marks on "uncertainty" or "lack of specific knowledge" to be able to verify those recommendations. However, that fraction of patterns was accepted with "surprisingly interesting" mark for further attention in their daily troubleshooting workarounds.

6 Conclusion and Future Work

We demonstrated ways to generate labeled data sets for training interpretable ML models that reveal rules and factors leading to unwanted states of KPIs of cloud applications. The goal of our project is to conduct RCA of KPI degradations with pretrained models while continuously updating it in a separate pipeline. Based on those models we can build troubleshooting and proactive support features (actionable recommender systems) that automatically alert on potential conditions/rules occurring in the application as an explanation of the KPI misbehavior.

In the modeled use case of application self-diagnostics, those conditions learned from the product usage and self-performance data can be shipped as AI-visible and customer-specific "rules" extracted based on granular time series data. Importance scoring of features is an alternative way to explain long-term behavior of the application and use it for performance optimization as well as situation-based troubleshooting purposes. We also discussed challenges with data labeling using a KPI as a source and techniques to overcome potential noise for training enough accurate ML models. Those models and research results were validated with product experts.

It is worth noting that in our current study we did not focus on the identification of the best/performant classification algorithm or model selection criteria as a primary objective, because of lack of human verified data sets and labeled data on KPI deviations, and, hence, unavailability of benchmarking opportunity. Instead, we discussed the KPI-diagnosis problem from different angles. On the other side, from productization/implementation perspectives, it is more viable and effective that the automated RCA recommender acts on high quality rules obtained using efficient induction algorithms handling noise, such as RIPPER [Cohen 1995]. This is a specific task in our future work plan. In terms of building more reliable RCA recommender systems, such an agenda includes working with other advanced algorithms (XGBoost [Chen and Guestrin 2016]) as well for higher accuracy tree boosting and domain-agnostic explainability frameworks (such as LIME [Ribeira et al 2016]) for local explainability upon availability of human inputs on incident instances on KPIs. Data imbalance might remain an issue in many application scenarios, therefore, alternative ways to overcome this problem need to be considered, e.g. oversampling framework SMOTE [Chawla et al 2011].

We also plan to validate our results further in multiple environments. In addition, it would be interesting to tackle the case when several KPIs need to be explained in various combinations in their behavior. Therefore, multi-labeling and relevant methods/algorithms might be appropriate to research on.

At this stage of our research, highly positive corporate reviews of our approaches (subject to a patented analytics) and models pave the path to productizations in VMware observability platforms ([VMware Aria 2022], [VMware Aria Logs 2023], [VMware Aria Nets 2023], [VMware Tanzu 2022].

Explainable AI is the next phase for many technologies to modernize their solutions for tomorrow's market requirements. We discussed such a challenge in realizing an effective management of large-scale cloud infrastructures and applications in terms of performance diagnosis. Automated RCA is a highly demanded solution in various closely related domains, such as cellular networks [Mdini 2019] and cloud databases [Ma et al 2020] where special and domain-specific modeling are adopted, which are not easily achievable in case of cloud infrastructures, thus leading the research towards more generic and self-supervised approaches.

Acknowledgements

The research is conducted within the ADVANCE Research Grants provided by the Foundation for Armenian Science and Technology. A. Poghosyan and T. Bunarjyan were funded by Science Committee of RA, in the frames of the research project № 20TTAT-AIa014.

References

[Baghdasaryan et al 2022] Baghdasaryan, A., Bunarjyan, T., Poghosyan, A., Harutyunyan, A., El-Zein, J.: On AI-driven customer support in cloud operations, Proc. 3nd Workshop on Collaborative Technologies and Data Science in Smart City Applications (CODASSCA 2022), Yerevan, Armenia, August 23-24, 32-35 (2022).

[Barredo Arrieta et al 2020] Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., and Herrera, F.: Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI, Information Fusion 58, 82-115 (2020).

[Bunarjyan et al 2020] Bunarjyan, T.A., Harutyunyan, A.N., Poghosyan, A. V., Han Vinck, A.J., Chen, Y., Hovhannisyan, N.A.: Estimating efficient sampling rates of metrics for training accurate machine learning models, Proc. 2nd Workshop on Collaborative Technologies and Data Science in Smart City Applications (CODASSCA 2020), Yerevan, Armenia, September 14-17, 143-152 (2020).

[Chawla et al 2011] Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P: SMOTE: Synthetic Minority Over-sampling Technique, https://arxiv.org/abs/1106.1813 (2011).

[Chen et al 2020] Chen, Z, Kang, Y., Li, L., Zhang, X., Zhang, H., Xu, H., Zhou, Y., Yang, L., Sun, J., Xu, Zh., Dang, Y., Gao, F., Zhao, P., Qiao, B., Lin, Q., Zhang, D., Lyu, M.: Towards intelligent incident management: why we need it and how we make it. Proc. 28th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering (ESEC/FSE 2020), 1487–1497 (2020).

[Chen and Guestrin 2016] Chen, T., Guestrin, C.: XGBoost: A scalable tree boosting system, https://arxiv.org/pdf/1603.02754.pdf (2016).

[Clark and Boswell 1998] Clark, P., Boswell, R.: Rule induction with cn2: Some recent improvements, Proc. European Working Session on Learning (1998).

[Cohen 1995] Cohen, W.: Fast effective rule induction, Proc. Twelfth International Conference on Machine Learning, Tahoe City, California, July 9–12, 115-123 (1995).

[Fürnkranz et al 2012] Fürnkranz, J., Gamberger, D., Lavrac, N.: Foundations of Rule Learning. Springer-Verlag (2012).

[Harutyunyan et al 2014] Harutyunyan, A.N., Poghosyan, A.N., Grigoryan, N.M, Marvasti, M.: Abnormality analysis of streamed log data, Proc. 2014 IEEE Network Operations and Management Symposium (NOMS 2014), Krakow, Poland, May 5-9, 1–7 (2014).

[Harutyunyan et al 2018] Harutyunyan, A.N., Poghosyan, A.V., Grigoryan, N.M., Kushmerick, N., Beybutyan, H.: Identifying changed or sick resources from logs, Proc. IEEE 3rd International Workshops on Foundations and Applications of Self* Systems (FAS*W), Trento, Italy, September 3-7, 86–91 (2018).

[Harutyunyan et al 2019] Harutyunyan, A.N., Poghosyan, A.V., Grigoryan, N.M., Hovhannisyan, N.A., Kushmerick, N.: On machine learning approaches for automated log management, Journal of Universal Computer Science 25(8), 925–945 (2019).

[Harutyunyan et al 2020(1)] Harutyunyan, A.N., Grigoryan, N.M., Poghosyan, A.V., Dua, S, Antonyan, H., Aghajanyan, K., Zhang, B.: Intelligent troubleshooting in data centers with mining evidence of performance problems, Proc. 2nd Workshop on Collaborative Technologies and Data Science in Smart City Applications (CODASSCA 2020), Yerevan, Armenia, September 14-17, 169-180 (2020).

[Harutyunyan et al 2020(2)] Harutyunyan, A.N., Grigoryan, N.M., Poghosyan, A.V.: Fingerprinting data center problems with association rules, Proc. 2nd Workshop on Collaborative Technologies and Data Science in Smart City Applications (CODASSCA 2020), Yerevan, Armenia, September 14-17, 159-168 (2020).

[Harutyunyan et al 2022] Harutyunyan, A.N., Aghajanyan, N.K., Harutyunyan, L.A., Poghosyan, A.V., Bunarjyan, T.A., Han Vinck, A.J.: On diagnosing cloud applications with explainable AI, Proc. 3nd Workshop on Collaborative Technologies and Data Science in Smart City Applications (CODASSCA 2022), Yerevan, Armenia, August 23-24, 23-26 (2022).

[HPE 2022] HPE Systems Insight Manager: www.support.hpe.com/hpesc/public/docDisplay?do-cId=c05330372 (2022).

[Jolliffe and Cadima 2016] Jolliffe, I.T., Cadima, J.: Principal component analysis: a review and recent developments, Philosophical transactions. Series A, Mathematical, physical, and engineering sciences (2016).

[Josefsson 2017] Josefsson, T.: Root-cause analysis through machine learning in the cloud, Master Thesis. Uppsala University: https://uu.diva-portal.org/smash/get/diva2:1178780/FULL-TEXT01.pdf (2017).

[Konotenko 1994] Konotenko, I: Estimating attributes: analysis and extensions of RELIEF, Proc. European Conference on Machine Learning, Catania, Italy, April 6-8, 171-182 (1994).

[Lyu and Su 2023] Lyu, M.R., Su, Y.: Intelligent Software Engineering for Reliable Cloud Operations. In: Wang, L., Pattabiraman, K., Di Martino, C., Athreya, A., Bagchi, S. (eds) System Dependability and Analytics. Springer Series in Reliability Engineering (2023).

[Lyu et al 2021] Lyu, Y., Rajbahandur, G.K., Lin, D., Chen, B., Jiang, Z.M., Towards a consistent interpretation of AIOps models, ACM Transactions on Software Engineering and Methodology 31(1), 1–38 (2021).

[Mdini 2019] Mdini, M: Anomaly Detection and root cause diagnosis in cellular networks. PhD Thesis. Rennes (2019).

[Ma et al 2020] Ma, M, Yin, Zh, Zhang, Sh., Wang, Sh., Zeng, Ch., Jiang, X., Hu, H., Luo, Ch.: Diagnosing root causes of intermittent slow queries in cloud databases. PVLDB, 13(8): 1176□1189, 2020.

[Orange 2023] Orange Software: https://orangedatamining.com/.

[Poghosyan et al 2020(1)] Poghosyan, A.V., Harutyunyan, A.N., Grigoryan, N.M., Kushmerick, N.: Learning data center incidents for automated root cause analysis, Proc. 2nd Workshop on Collaborative Technologies and Data Science in Smart City Applications (CODASSCA 2020), Yerevan, Armenia, September 14-17, 181-190 (2020).

[Poghosyan et al 2020(2)] Poghosyan, A.V., Harutyunyan, A.N., Grigoryan, N.M., Pang, C., Oganesyan, G., Ghazaryan, S., Hovhannisyan, N.: W-TSF: Time series forecasting with deep learning for cloud applications, Proc. 2nd Workshop on Collaborative Technologies and Data Science in Smart City Applications (CODASSCA 2020), Yerevan, Armenia, September 14-17, 152-158 (2020).

[Poghosyan et al 2021(1)] Poghosyan, A.V., Harutyunyan, A.N., Grigoryan, N.M., Kushmerick, N: Incident management for explainable and automated root cause analysis in cloud data centers, Journal of Universal Computer Science 27(11), 1152-1173 (2021).

[Poghosyan et al 2021(2)] Poghosyan, A.V., Harutyunyan, A.N., Grigoryan, N.M., Pang, C., Oganesyan, G., Ghazaryan, S., Hovhannisyan N.: An enterprise time series forecasting system for cloud applications using transfer learning, Sensors 21(5:1590), 1-28 (2021).

[Poghosyan et al 2022] Poghosyan, A.V., Harutyunyan, A.N., Grigoryan, N.M., Pang, C.: Root cause analysis of application performance degradations via distributed tracing, Proc. 3nd Workshop on Collaborative Technologies and Data Science in Smart City Applications (CODASSCA 2022), Yerevan, Armenia, August 23-24, 27-31 (2022).

[Ribeira et al 2016] Ribeira, M.T., Singh, S., Guestrin, C.: Why should I trust you?: Explaining the predictions of any classifier. arXiv: 1602.04938v1 (2016).

[SDDC 2017] Self-driving cars to self-driving data centers: https://www.broadcom.com/sw-tech-blogs/mainframe/self-driving-cars-self-driving-data-centers (2017).

[Self-Monitoring 2022] Self-Monitoring Metrics for vRealize Operations Manager. https://docs.vmware.com/en/vRealize-Operations/8.6/com.vmware.vcom.metrics.doc/GUID-A286313E-BC21-4965-8850-4A6E4D8E4459 (2022).

[Sole et al 2017] Sole, M., Muntes-Mulero, V., Rana, A.I., and Estrada, G.: Survey on models and techniques for root-cause analysis. arXiv: 1701.08556v2, (2017).

[VMware Aria 2022] VMware Aria Operations: https://www.vmware.com/products/vrealize-operations (2022).

[VMware Aria Logs 2023] VMware Aria Operations for Logs, https://www.vmware.com-/latam/products/vrealize-log-insight (2023).

[VMware Aria Nets 2023] VMware Aria Operations for Networks, https://www.vmware.com-/latam/products/vrealize-network-insight (2023).

[VMware Prods 2023] VMware products, https://www.vmware.com/products (2023).

[VMware Skyline 2022] VMware Skyline: https://www.vmware.com/support/services/skyline (2022).

[VMware Tanzu 2022] VMware Tanzu Observability by Wavefront, https://tanzu.vmware.com-/observability (2022).

[Wang et al 2023] Wang, W., Chen, J., Yang, L., Zhang, H., Wang, Z.: Understanding and predicting incident mitigation time, Information and Software Technology 155 (2023).

[Yu and Liu 2003] Yu, L., Liu, H.: Feature selection for high-dimensional data: A fast correlation-based filter solution, Proc. 20th International Conference on Machine Learning, 2:856–863 (2003).

Journal of Universal Computer Science, vol. 29, no. 11 (2023), 1319-1335
submitted: 19/5/2023, accepted: 30/9/2023, appeared: 28/11/2023 CC BY-ND 4.0

Deep Random Forest and AraBert for Hate Speech Detection from Arabic Tweets

Kheir Eddine Daouadi
(Laboratory of Vision and Artificial Intelligence (LAVIA), Echahid Cheikh Larbi Tebessi University, Tebessa, Algeria
https://orcid.org/0000-0003-2348-7192, kheireddine.daouadi@univ-tebessa.dz)

Yaakoub Boualleg
(Laboratory of Vision and Artificial Intelligence (LAVIA), Echahid Cheikh Larbi Tebessi University, Tebessa, Algeria
https://orcid.org/0000-0001-9665-5594, yaakoub.boualleg@univ-tebessa.dz)

Oussama Guehairia
(Laboratory of LESIA, Mohamed Khider University of Biskra, Biskra, Algeria
https://orcid.org/0000-0002-6755-4329, oussama.guehairia@univ-biskra.dz)

Abstract: Nowadays, hate speech detection from Arabic tweets attracts the attention of many researchers. Numerous systems and techniques have been proposed to address this classification challenge. Nonetheless, three major limits persist: the use of deep learning models with an excess of hyperparameters, the reliance on hand-crafted features, and the requirement for a huge amount of training data to achieve satisfactory performance. In this study, we propose Contextual Deep Random Forest (CDRF), a hate speech detection approach that combines contextual embedding and Deep Random Forest. From the experimental findings, the Arabic contextual embedding model proves to be highly effective in hate speech detection, outperforming the static embedding models. Additionally, we prove that the proposed CDRF significantly enhances the performance of Arabic hate speech classification.

Keywords: Twitter, Hate Speech Detection, Arabic Tweet Classification, Contextual Deep Random Forest, Fine-tuning, Pre-trained Contextual Embedding
Categories: H.3.1, H.3.2, H.3.3, H.3.7, H.5.1
DOI: 10.3897/jucs.112604

1 Introduction

Today, social media like Twitter have become a major part of our day-to-day life. This growing phenomenon has been very widespread in Arabic communities. Arabic Twitter users generate over 27 million tweets per day. The coming of Twitter has encouraged numerous axis research directions such as hate speech detection [Al-Hassan and Al-Dossari 2022, Duwairi et al. 2021, Mubarak et al. 2020], sentiment classification [Chamlertwat et al. 2012, Kalampokis et al. 2016, Baker et al. 2020, Abirami and Askarunisa 2016], bot detection [Daouadi et al. 2019b, Daouadi et al. 2020], organization detection [Daouadi et al. 2018a, Daouadi et al. 2019a, Daouadi et al. 2018b] and many more [Choi et al. 2013, Kalampokis et al. 2017] etc.

Recently, researchers have proposed many systems and techniques for Arabic hate speech detection. However, three of the major limits confronted in this context owing to the leverage of deep learning models based on many parameters, the leverage of handcrafted features, and their accuracy performance are limited. Automatic hate speech detection from Arabic tweets using traditional machine learning algorithms such as Support Vector Machine (SVM) [Husain 2020b] and Naïve Bayes (NB) [Mulki et al. 2019] have shown good results. Nevertheless, they focused mainly on hand-crafted features like Term Frequency (TF), Term Frequency-Inverse Document Frequency (TF-IDF), and Bag of Word (BoW). On the other hand, the latest developments in the field of deep learning classifiers like Convolution Neural Network (CNN) [Alsafari et al. 2020a] and Bidirectional Encoder Representation for Transformer (BERT) [Mubarak et al. 2020] have already demonstrated a remarkable performance for hate speech detection from Arabic tweets.

In this paper, we present a thorough investigation of DRF and contextual embedding for hate speech detection from Arabic tweets. Contextual embedding aims to convert tweet words into numerical values by using the available embedding models. The DRF consists of two major parts: Multi-Grained Scanning (MGS) and Cascade Forest (CF). The MGS aims to handle the different sliding windows and inputs them to the decision tree algorithms to create a feature vector for different sliding window sizes. The CF enables feature processing in a layer-by-layer fashion, based on an ensemble of random forests. Each cascade layer receives data produced by its predecessor. The output of the CF is the predicted tweet class. Therefore, we illustrate the leverage of our proposed approach by classifying a recent benchmark dataset of Arabic hate speech into 5 distinct classes: general hate, religious, sexism, racial, or none. Moreover, we conducted an extensive parameter sensitivity to tune CDRF for our Arabic hate speech classification task. A comparison between CDRF and existing hate speech classification approaches is presented. Experimental results demonstrated that CDRF improves the accuracy of hate speech detection from Arabic tweets, and outperforms existing works with a considerable margin. Our proposition may offer an important advantage by using ensemble learning that achieves better performance results than one single classifier, while CDRF is simple and easy to train based only on a few parameters. The contributions of this paper are presented as follows:

- We proposed a novel and efficient deep learning architecture combining contextual embedding and Deep Forest. The suggested architecture improves the accuracy of hate speech detection from Arabic tweets and outperforms the latest state-of-the-art results.
- We compare the performance of the static embedding models (AraVec, ArWord2vec, Mazajak, and Fastext) as well as contextual embedding models like (AraBERT, AraElectra, Multilingual BERT, and XLM-Roberta). Experiments show that the Arabic contextual embedding models trained on Twitter data improve the accuracy results of hate speech detection from Arabic tweets.
- We present an empirical study to select the hyperparameters for the CDRF architecture for hate speech detection from Arabic tweets.

The remainder of this paper is structured as follows. Section 2 discusses state-of-the-art approaches. Section 3 details our proposed approach. Section 4 presents the experimental results, and Section 5 summarizes the manuscript.

2 Related Works

Today, hate speech detection from Arabic tweets has attracted the attention of many researchers in the world. Different systems have been proposed to face this social challenge. They leverage two main approaches: a deep learning approach and a traditional approach.

2.1 Traditional Approaches

Traditional classification approaches rely on feature engineering, which converts texts to feature vectors and classifies them with machine learning algorithms like SVM and NB. The majority of the features were lexical-based features such as TF-IDF, N-grams, and BOW. The most relevant traditional approaches are described as follows.

Authors in [Husain 2020b] investigate the impact of the preprocessing step on hate speech and offensive language classification. They showed that an intensive preprocessing technique demonstrates its significant impact on the detection rate. The best accuracy results are obtained using BoW and SVM, which yielded an F1 score result of 89% and 95% for offensive language and hate speech.

In a similar, in [Husain 2020a] the authors classify tweets as offensive or not based on BoW and TF-IDF features. They showed that the ensemble learning classifier (Bagging) outperformed the single learner classifier, which yielded an F1 score of 88%.

Besides, authors in [Chowdhury et al. 2020] highlight the importance of multiple platform datasets for the generalization of classifier performance of offensive language detection. Their experiment with TF-IDF and SVM yielded an F1 score result of 84%.

Likewise, authors in [Aljarah et al. 2021] use BoW, TF-IDF, TF, Profile, and emotion features to classify tweets as being hate speech or not. The best accuracy results are obtained using Random Forest, which yielded an F1 score result of 91.3%.

Furthermore, authors in [Mulki et al. 2019] reported the classification rate using SVM and NB with unigrams, bigrams, and trigrams. The best experimental results were obtained using NB, which yielded an F1 score of 74.4% for (Non-Hate vs. hate vs. abusive) and 89.6% for (Abusive and Hate vs. Non-Hate).

In a different strategy, in [Abozinadah and Jones Jr 2017] the authors leverage profile-based, Social Graph feature, and tweet-based features to distinguish abusive Twitter accounts from those non-abusive. They reported that the NB algorithm yielded the best accuracy results of 85%.

2.2 Deep Learning Approaches

Deep Learning (DL) approaches rely on neural networks, which automatically learn the representation of input texts with different levels of abstraction and use the acquired knowledge for the classification task. The majority of used embedding models are AraVec, Mazajk, and AraBERT. The most popular deep learning architectures adopted in the field of hate speech detection from Arabic tweets are CNN, BERT, Long Short Term Memory (LSTM), and Gated Recurrent Unit (GRU). Some examples of DL approaches are described as follows.

Authors in [Faris et al. 2020] use the AraVec model based on N-grams and Skip-gram to classify tweets into those hateful and Non-hate. The Hybrid CNN-LSTM is used for classification, which yielded an F1 score result of 71.69%.

In similar, authors in [Al-Hassan and Al-Dossari 2022] classify tweets into five classes: general hate, sexism, racial, religious, or none. They randomly initialized the word embedding and learn the embedding of each word using the training dataset. The hybrid CNN-LSTM achieved the highest accuracy results, which yielded an average F1 score result of 73%.

Besides, authors in [Alsafari et al. 2020a] investigate the impact of both neural network architectures and word embedding models on the accuracy rate. They train different embedding models based on an Arabic text corpus. Additionally, they compared different neural networks for each classification task. The highest accuracy results are achieved using the Skip-gram embedding model and CNN, which yielded F1 score results of 87.22%, 75.16%, and 70.80% for (Non-hate, Offensive, or Hate), (Non-Hate, Offensive, or Hate) and (Non-Hate, Offensive, Religion Hate, Gender Hate, ethnicity hate or Nationality hate) classification task, respectively.

In a similar study [Alghanmi et al. 2020], the authors use both contextual (AraBERT) and static (AraVec) embedding to classify tweets as Non-Hate, hateful, or abusive. The CNN achieved the best accuracy results and yielded an average F1 score of 72.1%.

Likewise, authors in [Alsafari et al. 2020c] use contextual Multilingual BERT embedding model with CNN, which yielded F1 score results of 87.03%, 78.99%, and 75.51% for (Non-hate vs. Offensive and Hate), (Non-Hate, Offensive or Hate) and (Non-Hate, Offensive, Religion Hate, Gender Hate, ethnicity hate or Nationality hate) classification task, respectively.

Furthermore, authors in [Haddad et al. 2020] use bidirectional GRU augmented with attention layer and AraVec embedding model to detect hate speech and offensive language tweets. Additionally, they investigate the impact of various pre-processing and oversampling techniques to increase the performance rate. They achieved F1 score results of 85.9% and 75% F1 score for offensive language and hate speech detection, respectively.

In different strategies, the authors in [Mubarak et al. 2020, Shapiro et al. 2022, Khezzar et al. 2023, Boulouard et al. 2022] use transfer learning based on different Arabic Bert models, the proposed approaches achieved promising results for classifying Arabic hate speech.

2.3 Gaps and Contributions

To date, no study has proposed a DL technique based on non-neural networks. In this paper, we investigated the use of non-neural networks for hate speech detection from Arabic tweets.

Our choice goes somewhat in contradiction to that of many DL and traditional learning techniques, which are based on huge amounts of hyperparameters, while their performance is limited. Besides, their computational efficiency is suboptimal, and their interpretability is challenging.

3 Methodology

To overcome the previously mentioned challenges, we proposed a special DL architecture known as Contextual Deep Random Forest (CDRF). As highlighted in Figure 1, our proposition consists of 4 major steps. Specifically, Pre-processing, Contextual Embedding, Multi-Grained Scanning, and Cascade Forest are described as follows.

3.1 Pre-processing

The inputs of our proposition are the textual content of tweets, which consists of raw tweet content as highlighted in Figure 1. This step is dedicated to cleaning the textual content of tweets to produce more consistent and standard tweets. We performed some pre-processing tasks as follows [Al-Hassan and Al-Dossari 2022]:

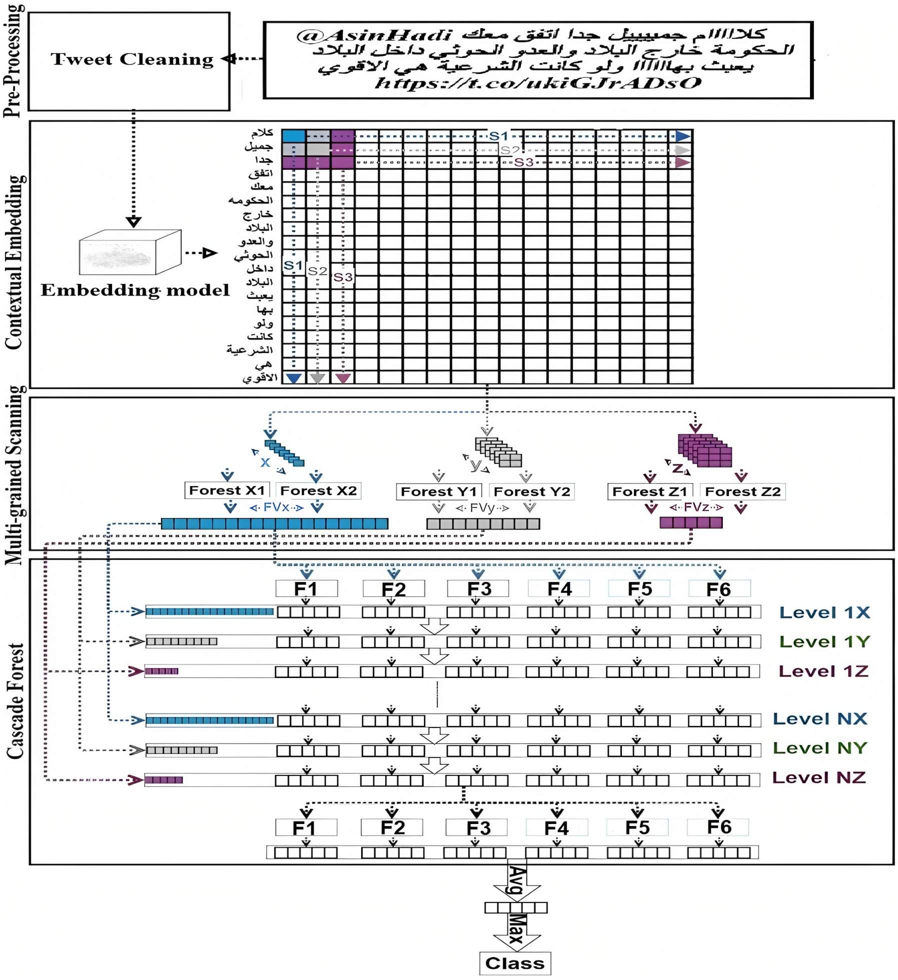


Figure 1: Our proposed CDRF approach.

- Removing Tweet features: the RT word, user mentions '@', hashtag symbol'#' URLs, punctuation, numerical characters, and special characters.
- non-Arabic letters, new lines as well as diacritics.
- Removing repeated letters: such as (مرحباااا) which means “Helloooooo”, to be (مرحبا), which means “Hello”.
- Arabic letters normalization: in Arabic language, they are different variations for some letters' representation which are:
  - Letter (Taa Marbouta) (ة) which can be mistaken and written as (ه), we normalized it to (ه).
  - Letter (Alef) (أ) that has the forms (ا, آ, إ, and أ), all the these letters are standardized into (ا).

- letter Dash that is leveraged to expand the word (مرحبا) to (مر—حباااا) has been removed.
- Letter (Alef Maqsora) (ي) has been normalized to (ى).

3.2 Contextual Embedding

The output of the previous steps is the normalized tweet content. Pre-trained contextualized word embedding plays an important role in constructing contextual tweet representation. The contextual relationships extracted by the embedding model are useful for detecting relevant contextual similarities. Contextual embeddings are calculated through a process involving :

- Tokenization: The input text (like a tweet) is divided into tokens (individual words or subwords). Each token is given an index, and special tokens like [CLS] and [SEP] might be added to the structure.
- Word Embedding Lookup: Each token's index is used to retrieve its pre-trained word embedding from a matrix. This matrix contains vectors representing a wide range of words.
- Positional Embeddings: Since the models don't naturally understand word order, positional embeddings are added to each token's embedding. These embeddings indicate where the token is located in the text.
- Attention Mechanism: Contextual embeddings use attention, allowing each token to consider information from all other tokens. Attention scores determine how important each token's connection is with others.
- Multiple Layers and Attention Heads: Contextual embeddings are produced through multiple layers and attention heads. Each layer refines embeddings by incorporating various contextual levels. Attention heads capture different relationships between words.
- Transformer Architecture: Models like BERT and GPT use a transformer architecture, composed of encoder and decoder layers. BERT uses only the encoder to understand input context. BERT is bidirectional, considering both the left and right context. These capture complex context relationships.
- Contextualized Embeddings: Final embeddings are contextualized, representing a token's meaning within its context. Words with multiple meanings can have different embeddings based on context.

To train our CDRF, all tweet instances are normalized to the longest tweet in our data by applying zero embedding. The contextualized embedding is used as input for the following step.

3.3 Multi-Grained Scanning

The major step of DRF is Multi-Grained Scanning (MGS), whose goal is to preserve the relationships between the word embedding vectors. The Sliding Window (SW) is the core building block MGS, which aims to handle the spatial relationships between the contextual Word Embedding vector. This helps to analyze the presence of such patterns or features in the tweet text. In this step, a sliding $S_j \in R^{k*k}$ is applied to the contextual tweet representation Emb_{Tweet}. As illustrated in Figure 1, where three different SW sizes are highlighted. The Stride (s) can shift the SW across s words for each step. The

Sliding operation produces N matrices instances (i.e. x, y, and z (cf Figure 1)) for each SW as follows:

$$N_i = E((ML - S_i)/s + 1) * E((dim - S_i/s + 1) \quad (1)$$

where ML is the maximum length of a tweet in our data, dim represents the dimension of WE, S_i represents the i^{th} sliding window size, s represent the stride, and E(z) represents the integer part of z.

Let MI_i be the Matrices Instance outputted by the i^{th} sliding window operation and NF represent the number of forests used in MGS. Each MGS forest (i.e. Forest X1, Forest X2, Forest Y1, Forest Y2, Forest Z1, and Forest Z2 (cf Figure 1)) outputs class probabilities as follows:

$$MGSForest_{nf}(MI_{ni}) = CP_{ni}^{nf} \quad (2)$$

Where $nf \in [1, NF]$, $ni \in [1, Number\ of\ instances]$, and $\mid CP_{ni}^{nf} \mid$ represent the number of classes. The output of MGS are the features vectors FV_{nsw} (i.e. FVx, FVy, and FVz (cf Figure 1)), which is the concatenation of all $(CP_{ni}^{nf})_{nsw}$ of the same SW S_{nsw} as described in the following equation:

$$FV_{nsw} = (CP_1^1)_{nsw} + ... + (CP_{ni}^{nf})_{nsw} \quad (3)$$

where $nsw \in [1, Number\ of\ Sliding\ Window]$ and :

$$\mid FV_{nsw} \mid = N_{nsw} * NF * \mid (CP_{ni}^{nf})_{nsw} \mid \quad (4)$$

3.4 Cascade Forest

As presented in Figure 1, the Cascade Forest (CF) has their own structure that performs layer-by-layer feature processing [Boualleg et al. 2019, Guehairia et al. 2020, Daouadi et al. 2021]. The output feature vectors from MGS are inputted through the cascade layers from the first to the last one. Each layer of the CF is an ensemble of random forests that receives data from the previous layer. Let NF' is the number of forests in CF (i.e. F1, F2, F3, F4, F5, and F6 (cf Figure 1)), for each layer (i.e. Level 1X … Level NZ (cf Figure 1)) a Class Distribution CD vector is produced as follows:

$$\mid CD_{nsw}^{nf'} \mid = NF\text{'} * number\_of\_classes \quad (5)$$

The final layer gets the vectors of class distribution probabilities from the previous cascade layer to calculate the tweet class as follows:

$$Probability\_classes = Avg(CD_{nsw}^{nf'}) \quad (6)$$

Finally, we used the majority voting technique to calculate the final predicted class label as follows:

$$Y = Argmax(Probability\_classes) \quad (7)$$

where Probability_classes is the vector of the probability distribution of each class.

To reduce the risk of overfitting, we performed K-fold cross-validation to calculate the class probability of each random forest. Notably, each tweet instance in the training data was leveraged K-1 times, which resulted in K-1 classes, which are then averaged to calculate the class probabilities as improved feature vectors for the next cascade level. The performance metrics of a whole level have been evaluated on the test sets after the expansion of a novel level. The training procedure ends where there is no significant gain in performance. In contrast, to the majority of DL models, whose complexity is fixed, CDRF chooses its complexity as terminating training procedure when adequate.

4 Experiments and Evaluation

In this section, we present the experimental and evaluation results we carried out to show the effectiveness of CDRF. Through the experiments on the benchmark Twitter dataset, we try to find answers to the following research questions:

- Can a contextual pre-trained embedding model improve the accuracy of hate speech detection from Arabic tweets?
- Can a Deep Learning architecture based on a non-neural network classify hate speech from Arabic tweets accurately?

To answer these questions, we first introduce the datasets used in our experiments. Subsequently, we provide an analysis of the performance results, followed by a comprehensive discussion of the findings.

	NH	**GH**	**ReH**	**RaH**	**S**	**Total**
Number of tweets	8332	1397	722	526	657	11634
Word count	96.9 K	17.2 K	9.2 K	6.6 K	8.3 K	138.3 K
Unique words	29 K	8.7 K	4.9 K	4 K	4.4 K	37.9 K
Avg words per tweet	11.6	12.3	12.7	12.7	12.6	11.9

Table 1: Overview of the ground truth after pre-processing (K denotes a thousand, NH = Non-Hate Speech, GH = General Hate Speech, ReH = Religious Hate Speech, RaH = Racial Hate Speech, S = Sexism).

4.1 Datasets

To evaluate our proposition, we refer to the dataset published in [Al-Hassan and Al-Dossari 2022]. An Overview of this dataset after pre-processing is shown in Table 1. For collecting the tweets, the authors used the Tweepy library with a list of hashtags that trigger and attract Twitter hateful content. Balancing the number of non-hate tweets and hate tweets does not reflect the real situation and is not realistic. The authors identified a list of hashtags that contains surely both non-hateful and hateful content to preserve the realistic and natural scenario. Then, the retrieved tweets were manually annotated by 2 annotators. The annotators were provided with a guide to distinguish the tweets classes, which yielded 11634 labeled tweets out of 37K collected tweets. As presented in Table 1, the minor set of tweets is the racial hate speech class, while the major set is the non-hate class.

In addition, we evaluate the proposed approach using train/test datasets published in [Alsafari et al. 2020c]. The dataset comprises 5340 tweets (3738 for training and 1602 for testing) manually classified as Clean, Offensive, Religious hate speech, Gender hate speech, Ethnicity hate speech, and Nationality hate speech. The annotation process was conducted by three native speakers from the Gulf region, including two females and one male, all possessing a higher level of education. These annotators were initially provided with task descriptions, guidelines, and three illustrative examples for each hate category.

4.2 Experimental Setup

To evaluate our proposition, we used the 10-fold cross-validation to estimate the performance metrics. The data was divided into ten equally sized segments while preserving the balance of each class in the original dataset. One of the segments was used as testing data and the other ones were used as training data. This is repeated ten times, the average F1-score metrics were obtained using the ten iterations. In our experiments, we used Scikit-Learn [Pedregosa et al. 2011], Ktrain [Maiya 2022], and Keras [1] which uses Tensorflow [2] as back-end. All the experiments were executed on a machine equipped with an Intel Core i7-7700 16GB RAM. The initial parameters of CDRF are SW=1*1, NT=50, NF=5, NT'=2, and NF'=100.

4.3 Evaluation Metrics

To evaluate the performance of our proposition, we used different evaluation metrics that can judge the model. Since the task of hate-speech detection is an imbalanced classification problem, we will pay big attention to the Macro-averaged and Weighted-averaged to calculate the overall performance measures described as follows.

Precision (P) (also called positive predictive value) represents the fraction of correctly classified positive observations over the total observations classified as positive. For instance, the Precision of the Non-Hate class is calculated as follows:

$$P_{Non-Hate} = CC_{Non-Hate}/TC_{Non-Hate} \tag{8}$$

where $CC_{Non-Hate}$ is the number of tweets correctly classified as Non-Hate and $TC_{Non-Hate}$ is the total number of tweets classified as Non-Hate.

Recall (R) is the fraction of correctly classified positive observations over the total positive observations. For instance, the Recall of the Non-Hate class is calculated as follows:

$$R_{Non-Hate} = CC_{Non-Hate}/TN_{Non-Hate} \tag{9}$$

where $CC_{Non-Hate}$ is the number of tweets correctly classified as Non-Hate and $TN_{Non-Hate}$ is the total number of Non-Hate tweets.

F1-score (F1) represents the harmonic mean between Recall and Precision. For instance, the $F1_N$ of the Non-Hate class is calculated as follows:

$$F1_{NH} = 2(P_{NH} * R_{NH})/(P_{NH} + R_{NH}) \tag{10}$$

[1] Keras is an open-source neural network library written in Python which is designed for fast experimentation with deep neural networks. https://keras.io

[2] TensorFlow is an open-source software library for data flow programming across a range of tasks, which is used for different machine learning applications, such as neural networks. https://www.tensorflow.org

4.4 Results

We performed extensive experiments on the proposed CDRF for hate speech detection from Arabic tweets. We compared the accuracy results of both static and contextual embedding models, namely: ArWord2vec CBOW-5 (W2V_C_5) and ArWord2vec SG-5 (W2V_S_5) [Fouad et al. 2020]; Mazajak 100M-SG (M_100_S) and Mazajak 100M-SG (M_250_S) [Farha and Magdy 2019]; AraVec Tweet-CBOW (AV_Twt_C) and AraVec Tweet-SG (AV_Twt_S) [Soliman et al. 2017]; Fasttext [Joulin et al. 2016]; AraBERT and AraBERTv02-twitter (AraB_v02_T) [Antoun et al. 2020a]; AraElectra [Antoun et al. 2020c]; Multilingual BERT (MBERT) [Devlin et al. 2018], XLM-Roberta (XLM-R) [Conneau et al. 2019]; BERT Arabic (BERTA) [Safaya et al. 2020], Multi-dialectal Arabic BERT (MdABERT) [Talafha et al. 2020]; and MARBERT [Abdul-Mageed et al. 2020].

The first experiment aims to select the word embedding model that attains the highest F1 score results. Table 2 shows a comparison of both static and contextual embedding models. The results show that the Arabic contextual embedding model trained on Twitter data outperforms the other ones with a considerable, yielded Macro-averaged F1-score result of 0.55 and a Weighted-averaged F1-score result of 0.77. The primary factors are the datasets used to train the embedding model and the effectiveness of contextual embeddings in capturing the meaning of words by analyzing the words that appear around them in sentences or texts.

The second set of experiments aims to select the optimal SW sizes. The accuracy rates of multiple and single sizes of sliding windows are presented in Table 3. The results show that the multiple sizes of SW increase the F1-score as they extract more N-gram features, yielding 0.58 and 0.78 for the Macro-averaged F1-score and the Weighted-averaged

Model	**NH**	**S**	**ReH**	**GH**	**RaH**	**M**	**W**
Random	75.81	41.54	15.75	39.12	21.66	38.78	63.29
M_100_S	87.04	45.75	56.89	31.25	32.96	50.78	73.69
M_250_S	86.78	43.73	61.82	29.03	26.16	49.51	73.12
W2V_C_5	86.58	45.68	58.13	34.37	27.68	50.49	73.57
W2V_S_5	86.22	42.20	58.68	33.16	25.49	49.15	72.90
AV_Twt_S	87.30	44.00	60.99	31.82	**33.58**	51.54	74.13
AV_Twt_C	87.17	43.19	61.43	31.40	32.05	51.05	73.85
Fasttext	86.45	44.07	57.28	34.06	26.75	49.72	73.26
MdABERT	87.46	48.28	51.00	**42.86**	25.83	51.09	74.84
BERTA	87.33	47.28	52.34	41.83	25.59	50.87	74.64
AraElectra	87.58	47.83	62.71	38.55	26.50	52.63	75.16
AraBERT	88.05	47.39	57.41	42.52	28.84	52.84	75.70
MARBERT	88.40	50.00	61.40	41.62	30.74	54.43	76.33
AraB_v02_T	**88.59**	**51.49**	**64.61**	41.32	31.26	**55.45**	**76.73**
MBERT	86.51	38.46	44.72	39.01	24.58	46.65	72.69
XLM-R	85.22	23.47	41.65	38.02	22.55	42.18	70.52

Table 2: Comparison of different embedding models based on F1-score (%) (M=Macro-Averaged, W=Weighted-Averaged).

F1-score, respectively

The third experiment aims to select the optimal number of NF. The effect of NF on the F1-score is presented in Table 4. The F1-score results increased when NF increased from 5 to 10. Thus, it decreased when NF increased further.

The fourth experiment aims to select the optimal number of NT. The F1-score results remain fixed when NT increased from 50 to 200 As shown in Table 5. Thus, it decreased when NT increased further.

The fifth experiment aims to find the optimal number of NF'. As highlighted in Table 6, the F1-score increased when NF' increased from 2 to 4. Thus, it decreased when NF' increased further.

The sixth experiment aims to select the optimal number of NT'. The F1-score increased when NT' increased from 100 to 300 as highlighted in Table 7. Thus kept stable when NT' increased further.

The best parameters of CDRF for our hate speech detection task are presented in Table 8 where NF (F)= 10, NT (T)= 200, SW = {1, 3, 5, 7, 9}, NF'(F') = 4 and NT'(T') = 300. The number of the level of Cascade Layer (C) is 2 when using the optimized parameters, while (N') is the count of nodes in the trees fixed to 10 nodes, and the total

SWS	NH	S	ReH	GH	RaH	M	W
1	88.59	51.49	64.61	41.32	31.26	55.45	76.73
3	88.71	46.42	61.57	45.14	31.63	54.69	76.82
5	88.64	52.09	63.58	41.93	31.63	55.57	76.83
7	88.76	49.40	58.74	45.66	32.15	54.94	76.93
9	88.72	48.62	59.70	45.94	32.52	55.10	76.97
1,3	88.68	46.59	69.56	44.70	34.44	56.79	77.38
3,5	88.56	45.81	69.62	44.25	33.33	56.32	77.15
5,7	88.72	47.02	70.20	44.76	35.14	57.17	77.51
7,9	88.70	46.60	70.07	44.65	34.64	56.93	77.43
1,3,5	88.77	47.69	69.26	45.35	35.69	57.35	77.62
3,5,7	88.84	47.88	69.57	45.24	**36.27**	57.56	77.71
5,7,9	88.86	48.24	**70.87**	44.83	35.79	57.72	77.76
1,3,5,7	88.94	49.70	69.22	45.90	36.00	57.95	77.94
3,5,7,9	88.71	46.95	70.06	44.78	34.81	57.06	77.48
1,3,5,7,9	**89.01**	**53.83**	64.92	**46.29**	36.22	**58.05**	**78.01**

Table 3: The effect of multiple sliding windows on F1-score (%).

NF	NH	S	ReH	GH	RaH	M	W
5	89.01	53.83	64.92	46.29	36.22	58.05	78.01
10	**89.43**	**58.51**	64.23	**46.45**	**37.15**	**59.15**	**78.59**
15	89.11	54.04	65.36	46.39	36.67	58.31	78.15
20	88.97	49.28	69.28	45.83	36.40	57.95	77.94
25	88.75	47.56	**69.34**	45.34	35.39	57.27	77.59

Table 4: The effect of NF on F1-score (%).

NT	NH	S	ReH	GH	RaH	M	W
50	89.43	58.51	64.23	46.45	37.15	59.15	78.59
100	89.42	58.17	66.14	47.98	39.16	60.17	78.96
150	89.47	58.88	66.71	**48.02**	39.76	60.57	79.10
200	**89.58**	**60.27**	**69.16**	47.26	**40.16**	**61.29**	**79.34**
250	89.45	58.27	65.05	46.47	37.35	59.32	78.66

Table 5: The effect of NT on F1-score (%).

NF'	NH	S	ReH	GH	RaH	M	W
2	89.58	60.27	69.16	47.26	40.16	61.29	79.34
4	**89.62**	**61.77**	72.90	46.09	**41.13**	**62.30**	**79.59**
6	89.54	61.43	**73.76**	45.43	40.64	62.16	79.46
8	89.42	58.69	66.54	**47.86**	39.27	60.36	79.01
10	89.52	60.02	67.70	47.65	39.72	60.92	79.22

Table 6: The effect of NF' on F1 score F1-score (%).

NT'	NH	S	ReH	GH	RaH	M	W
100	89.62	**61.77**	72.90	46.09	41.13	62.30	79.59
200	89.60	57.58	74.51	48.04	42.22	62.39	79.72
300	**89.84**	58.98	**75.31**	**48.98**	**43.63**	**63.35**	**80.20**
400	89.47	58.88	66.71	48.02	39.76	60.57	79.10
500	89.54	60.03	74.32	45.69	40.40	62.00	79.44

Table 7: The effect of NT' on F1-score (%).

Hyperparameters	Embedding Model	SW	NF	NT	NF'	NT'
Value	AraBERTv02-Twitter	{1,3, 5, 7, 9}	10	200	4	300

Table 8: Optimized parameters of CDRF.

of SW (N) used in CDRF is 5. The total Trainable Parameters in CDRF is estimated using the following equations:

$$TP_{CDRF} = TP_{MGS} + TP_{CF} \tag{11}$$

$$TP_{MGS} = (F * T * N) \tag{12}$$

$$TP_{CF} = (T' * F' * N * N' * C) + (T' * F' * N) \tag{13}$$

According to the aforementioned Equations, the number of TP in CDRF is equal to 0.152 million trainable parameters, unlike most of the state-of-the-art models that employ more than a million parameters.

Thereafter, the F1-score of CDRF is compared against four existing traditional machine learning approaches [Mulki et al. 2019, Husain 2020a, Chowdhury et al. 2020, Miller et al. 2017], six existing DL approaches, [Al-Hassan and Al-Dossari 2022, Alsafari

Approach	**NH**	**S**	**ReH**	**GH**	**RaH**	**M**	**W**
[Mulki et al. 2019]	0.85	0.21	0.10	0.12	0.09	0.27	0.64
[Husain 2020a]	0.86	0.42	0.50	0.25	0.24	0.45	0.71
[Chowdhury et al. 2020]	0.87	0.41	0.54	0.30	0.29	0.48	0.73
[Al-Hassan and Al-Dossari 2022]	0.86	0.43	0.59	0.31	0.25	0.49	0.73
[Alsafari et al. 2020c]	0.85	0.42	0.56	0.45	0.20	0.50	0.73
[Haddad et al. 2020]	0.87	0.50	0.64	0.49	0.27	0.55	0.76
[Alghanmi et al. 2020]	0.86	0.44	0.59	0.47	0.22	0.52	0.74
[Faris et al. 2020]	0.87	0.37	0.47	0.28	0.26	0.45	0.72
[Alsafari et al. 2020b]	0.87	0.49	0.56	0.47	0.24	0.53	0.75
[Mubarak et al. 2020]	0.88	0.49	0.67	0.41	0.37	0.57	0.77
[Khezzar et al. 2023]	0.89	0.55	0.65	**0.52**	0.42	0.61	0.79
Arabert + Random Forest	0.89	0.51	0.63	0.51	0.40	0.59	0.79
Arabert + [Miller et al. 2017]	0.89	0.47	0.61	0.55	0.37	0.58	0.79
Aragpt [Antoun et al. 2020b]	0.85	0.23	0.42	0.38	0.23	0.42	0.70
CDRF	**0.90**	**0.59**	**0.75**	0.49	**0.44**	**0.63**	**0.80**

Table 9: Comparison of CDRF and the latest baseline approaches based on F1-score.

et al. 2020c, Haddad et al. 2020, Alghanmi et al. 2020, Faris et al. 2020, Alsafari et al. 2020b], and three fine-tuning approaches [Mubarak et al. 2020, Khezzar et al. 2023, Antoun et al. 2020b]. While comparing CDRF with the baselines presented in Tables 9 and 10, the F1-score of CDRF is the highest. The main reasons are due to the contextual representation of the tweet. When compared with LSTM, BERT, and CNN, current DL hate speech classification models are built based on Neural Networks trained based on backpropagation. On the other hand, CDRF is a DL approach that leverages an ensemble of random forests and their training procedure doesn't depend on backpropagation. When compared with SVM, NB, and Bagging, CDRF has not extracted and designed hand-crafted features. In addition, CDRF deals with social media data, where the user uses many slangs, abbreviations, etc. (i.e. the language structure is not preserved). CDRF analyzes the sequence of words by correlating them with past and future words.

Even though quite successful, previous DL Arabic hate speech detection approaches are so expensive, as they are built upon Neural Networks trained based on backpropagation with so many hyperparameters. As an example, taking [Mubarak et al. 2020, Husain and Uzuner 2021] are recent DL models that used AraBERT (with over 100 million parameters).On the contrary, CDRF is a DL approach based on an ensemble of Random Forest, where the training procedure doesn't depend on backpropagation, while using fewer hyperparameters than the existing approaches. Furthermore, current traditional approaches used handcrafted features, which faced the curse of dimensionality and data sparseness. Conversely, CDRF can automatically detect features from textual content. Finally, the comparison between the accuracy rate of CDRF and 14 baselines is performed, and the obtained result highlights the efficiency of CDRF against existing approaches. Resulting promising weighted averaged F1 score result of (> 80%). In Table 2, we can notice that: the minor accuracy results go for the multilingual contextual embedding models; the medium results go for Arabic static embedding models and contextual em-

Approach	**C**	**O**	**R**	**G**	**N**	**E**	**M**	**W**
[Mulki et al. 2019]	0.89	0.49	0.47	0.52	0.56	0.69	0.61	0.77
[Husain 2020a]	0.89	0.54	0.48	0.55	0.57	0.70	0.62	0.78
[Chowdhury et al. 2020]	0.91	0.54	0.52	0.55	0.64	0.68	0.64	0.79
[Al-Hassan and Al-Dossari 2022]	0.87	0.57	0.71	0.64	0.68	0.69	0.69	0.80
[Alsafari et al. 2020c]	0.91	0.72	0.71	0.68	0.74	0.77	0.76	0.85
[Haddad et al. 2020]	0.86	0.56	0.70	0.63	0.67	0.70	0.69	0.79
[Alghanmi et al. 2020]	0.95	0.67	0.68	0.67	0.73	0.75	0.74	0.86
[Faris et al. 2020]	0.94	0.61	0.66	0.63	0.71	0.73	0.71	0.85
[Alsafari et al. 2020b]	0.96	0.74	0.73	0.76	0.76	0.79	0.79	0.89
[Mubarak et al. 2020]	0.95	0.77	0.71	0.80	0.79	0.80	0.80	0.89
[Khezzar et al. 2023]	0.96	0.80	0.75	0.74	0.80	0.79	0.81	0.90
Arabert + RF	0.92	0.84	0.77	0.73	**0.84**	0.80	0.82	0.88
Arabert + [Miller et al. 2017]	0.92	0.76	0.77	0.75	0.78	0.79	0.80	0.87
Aragpt [Antoun et al. 2020b]	0.88	0.47	0.45	0.48	0.56	0.67	0.58	0.76
CDRF	**0.96**	**0.85**	**0.80**	**0.82**	0.83	**0.83**	**0.85**	**0.91**

Table 10: Comparison of F1-score of CDRF and the latest baseline approaches based using train/test datasets published in [Alsafari et al. 2020c] (C = Clean, O=Offensive, R=Religious hate speech, G = Gender hate speech, N = Nationality hate speech, E = Ethnicity hate speech)

.

bedding trained on general data while the major results go for the Arabic contextual model trained on Twitter data. In Table 9, we can notice that the minor accuracy results go for the racial hate speech class, while the major results go for the non-hate class. This is due to the severe class label imbalance. Thus, future work should be harnessed to face the problem of imbalanced data.

5 Conclusion

The proposed CDRF approach combines contextual embedding with Deep Random Forest, yet achieves the highest F1 results (> 80%). Our proposed approach highlights a new research area for hate speech detection approaches. Experiments show that: (1) the Arabic contextual embedding models trained on Twitter data can be effectively leveraged for hate speech detection and outperform multilingual contextual embedding models, static embedding model, and Arabic contextual models trained on non-Twitter data; (2) the proposed CDRF improve the accuracy of hate speech detection of Arabic tweets and outperform the latest state-of-the-art results. In future works, we plan to pursue several directions. First, we want to focus on the contextual embedding model and try to adjust their vocabulary to support the hate speech detection task. Second, we plan to test and compare different data augmentation techniques to overcome the challenge of imbalanced learning. From a research perspective, we plan to use CDRF for analyzing Arabic Tweets in some contexts to study the extent of hate speech conversations with public discussion.

References

[Abdul-Mageed et al. 2020] Abdul-Mageed, M., Elmadany, A., Nagoudi, E. M. B.: ”ARBERT MARBERT: deep bidirectional transformers for Arabic”. arXiv preprint arXiv:2101.01785., (2020).

[Abirami and Askarunisa 2016] Abirami, A. M., Askarunisa, A.: ”Feature Based Sentiment Analysis for Service Reviews”; J. Univers. Comput. Sci., 22(5), 650-670, (2016).

[Abozinadah and Jones Jr 2017] Abozinadah, E. A., Jones Jr, J. H.: ”A statistical learning approach to detect abusive twitter accounts”. In Proceedings of the international conference on compute and data analysis, 6-13, (May 2017).

[ALBayari et al. 2021] ALBayari, R., Abdullah, S., Salloum, S. A.: ”Cyberbullying Classification Methods for Arabic: A Systematic Review”; In The International Conference on Artificial Intelligence and Computer Vision, 375-385, (May 2021).

[Alghanmi et al. 2020] Alghanmi, I., Anke, L. E., Schockaert, S.: ”Combining BERT with static word embeddings for categorizing social media”; In Proceedings of the sixth workshop on noisy user-generated text, 28-33, (Nov 2020).

[Aljarah et al. 2021] Aljarah, I., Habib, M., Hijazi, N., Faris, H., Qaddoura, R., Hammo, B., Abushariah,M., Alfawareh, M.:”Intelligent detection of hate speech in Arabic social network: A machine learning approach”; Journal of Information Science, 47(4), 483-501, (2021).

[Alsafari et al. 2020a] Alsafari, S., Sadaoui, S., Mouhoub, M.: ”Effect of word embedding models on hate and offensive speech detection”; arXiv preprint arXiv:2012.07534., (2020).

[Alsafari et al. 2020b] Alsafari, S., Sadaoui, S., Mouhoub, M.: ”Deep Learning Ensembles for Hate Speech Detection”; In IEEE 32nd International Conference on Tools with Artificial Intelligence, 526-531, (2020).

[Alsafari et al. 2020c] Alsafari, S., Sadaoui, S., Mouhoub, M.: ”Hate and offensive speech detection on arabic social media”; Online Social Networks and Media, 19, 100096, (2020).

[Al-Hassan and Al-Dossari 2022] Al-Hassan, A., Al-Dossari, H.: ”Detection of hate speech in Arabic tweets using deep learning”; Multimedia systems, 28(6), 1963-1974,(2022).

[Antoun et al. 2020a] Antoun, W., Baly, F., Hajj, H.: ”Arabert: Transformer-based model for arabic language understanding”; arXiv preprint arXiv:2003.00104., (2020).

[Antoun et al. 2020c] Antoun, W., Baly, F., Hajj, H.: ”AraELECTRA: Pre-training text discriminators for Arabic language understanding”; arXiv preprint arXiv:2012.15516. (2020).

[Antoun et al. 2020b] Antoun, W., Baly, F., Hajj, H.: ”AraGPT2: Pre-trained transformer for Arabic language generation”; arXiv preprint arXiv:2012.15520., (2020).

[Baker et al. 2020] Baker, Q. B., Shatnawi, F., Rawashdeh, S., Al-Smadi, M., Jararweh, Y.: ”Detecting epidemic diseases using sentiment analysis of arabic tweets”. J. Univers. Comput. Sci., 26(1), 50-70, (2020).

[Boualleg et al. 2019] Boualleg, Y., Farah, M., Farah, I. R. (2019). Remote sensing scene classification using convolutional features and deep forest classifier. IEEE Geoscience and Remote Sensing Letters, 16(12), 1944-1948, (2019).

[Boulouard et al. 2022] Boulouard, Z., Ouaissa, M., Ouaissa, M., Krichen, M., Almutiq, M., Gasmi, K.: ”Detecting Hateful and Offensive Speech in Arabic Social Media Using Transfer Learning”; Applied Sciences, 12(24), 12823, (2022).

[Chamlertwat et al. 2012] Chamlertwat, W., Bhattarakosol, P., Rungkasiri, T., Haruechaiyasak, C.: ”Discovering Consumer Insight from Twitter via Sentiment Analysis”; J. Univers. Comput. Sci., 18(8), 973-992, (2012).

[Choi et al. 2013] Choi, D., Kim, J., Piao, X., Kim, P.: ”Text Analysis for Monitoring Personal Information Leakage on Twitter”; J. Univers. Comput. Sci., 19(16), 2472-2485, (2013).

[Chowdhury et al. 2020] Chowdhury, S. A., Mubarak, H., Abdelali, A., Jung, S. G., Jansen, B. J., Salminen, J.: ”A multi-platform Arabic news comment dataset for offensive language detection”; In Proceedings of the 12th Language Resources and Evaluation Conference, 6203-6212, (2020).

[Conneau et al. 2019] Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Edouard Grave, Ott,E. M.Zettlemoyer, M. Stoyanov, V.: ”Unsupervised cross-lingual representation learning at scale”; arXiv preprint arXiv:1911.02116., (2019).

[Daouadi et al. 2018a] Daouadi, K. E., Zghal Rebaï, R., Amous, I.: ”Towards a Statistical Approach for User Classification in Twitter”; In Machine Learning for Networking: First International Conference, 33-43,(Nov 2018).

[Daouadi et al. 2018b] Daouadi, K. E., Rebaï, R. Z., Amous, I.: ”Organization vs. Individual: Twitter User classification”; In International Workshop on Language Processing and Knowledge Management, (2018).

[Daouadi et al. 2019a] Daouadi, K. E., Rebaï, R. Z., Amous, I.: ”Organization, Bot, or Human: Towards an Efficient Twitter User Classification”; Computación y Sistemas, 23(2), 273-280, (2019).

[Daouadi et al. 2019b] Daouadi, K. E., Rebaï, R. Z., Amous, I.: ”Bot Detection on Online Social Networks Using Deep Forest”; In Proc. Int. Conf on Computer Science On-line Conference, 307-315, (Apr 2019).

[Daouadi et al. 2020] Daouadi, K. E., Rebaï, R. Z., Amous, I.: ”Real-Time Bot Detection from Twitter Using the Twitterbot+ Framework”; J. Univers. Comput. Sci., 26(4), 496-507, (2020).

[Daouadi et al. 2021] Daouadi, K. E., Rebaï, R. Z., Amous, I.: ”Optimizing semantic deep forest for tweet topic classification”; Information Systems, 101, 101801, (2021).

[Devlin et al. 2018] Devlin, J., Chang, M. W., Lee, K., Toutanova, K.: ”Bert: Pre-training of deep bidirectional transformers for language understanding”; arXiv preprint arXiv:1810.04805., (2018).

[Duwairi et al. 2021] Duwairi, R., Hayajneh, A., Quwaider, M.: ”A deep learning framework for automatic detection of hate speech embedded in Arabic tweets”; Arabian Journal for Science and Engineering, 46, 4001-4014, (2021).

[Farha and Magdy 2019] Farha, I. A., Magdy, W.: ”Mazajak: An online Arabic sentiment analyser”; In Proceedings of the fourth arabic natural language processing workshop, 192-198, (Aug 2019).

[Faris et al. 2020] Faris, H., Aljarah, I., Habib, M., Castillo, P. A.: ”Hate Speech Detection using Word Embedding and Deep Learning in the Arabic Language Context”; In Proceedings of the 9th International Conference on Pattern Recognition Applications and Methods, 453-460, (Feb 2020).

[Fouad et al. 2020] Fouad, M. M., Mahany, A., Aljohani, N., Abbasi, R. A., Hassan, S. U.: ”ArWordVec: efficient word embedding models for Arabic tweets”; Soft Computing, 24, 8061-8068, (2020).

[Guehairia et al. 2020] Guehairia, O., Ouamane, A., Dornaika, F., Taleb-Ahmed, A. (2020). Feature fusion via Deep Random Forest for facial age estimation. Neural Networks, 130, 238-252.

[Haddad et al. 2020] Haddad, B., Orabe, Z., Al-Abood, A., Ghneim, N.: ”Arabic offensive language detection with attention-based deep neural networks”; In Proceedings of the 4th workshop on open-source Arabic corpora and processing tools, with a shared task on offensive language detection, 76-81,(May 2020).

[Husain 2020a] Husain,F.: ”Arabic offensive language detection using machine learning and ensemble machine learning approaches”; arXiv preprint arXiv:2005.08946., (2020).

[Husain 2020b] Husain, F.: ”OSACT4 Shared Task on Offensive Language Detection: Intensive Preprocessing-Based Approach”; arXiv preprint arXiv:2005.07297, (2020).

[Husain and Uzuner 2021] Husain, F., Uzuner,O.: ”Transfer Learning Approach for Arabic Offensive Language Detection System–BERT-Based Model”. arXiv preprint arXiv:2102.05708, (2021).

[Joulin et al. 2016] Joulin, A., Grave, E., Bojanowski, P., Mikolov, T.: ”Bag of tricks for efficient text classification”; arXiv preprint arXiv:1607.01759. (2016).

[Kalampokis et al. 2016] Kalampokis, E., Karamanou, A., Tambouris, E., Tarabanis, K. A.: ”Applying Brand Equity Theory to Understand Consumer Opinion in Social Media”. J. Univers. Comput. Sci., 22(5), 709-734, (2016).

[Kalampokis et al. 2017] Kalampokis, E., Karamanou, A., Tambouris, E., Tarabanis, K.: ”On Predicting Election Results using Twitter and Linked Open Data: The Case of the UK 2010 Election”; Journal of Universal Computer Science, 23(3), 280-303, (2017).

[Khezzar et al. 2023] Khezzar, R., Moursi, A., Al Aghbari, Z.: ”arHateDetector: detection of hate speech from standard and dialectal Arabic Tweets”; Discover Internet of Things, 3(1), 1, (2023).

[Maiya 2022] Maiya, A. S.; ”ktrain: A low-code library for augmented machine learning”; The Journal of Machine Learning Research, 23(1), 7070-7075, (2022).

[Miller et al. 2017] Miller, K., Hettinger, C., Humpherys, J., Jarvis, T., Kartchner, D.: ”Forward thinking: Building deep random forests”; arXiv preprint arXiv:1705.07366., (2017).

[Mubarak et al. 2020] Mubarak, H., Rashed, A., Darwish, K., Samih, Y., Abdelali, A.: ”Arabic offensive language on twitter: Analysis and experiments”; arXiv preprint arXiv:2004.02192., (2020).

[Mubarak et al. 2020] Mubarak, H., Darwish, K., Magdy, W., Elsayed, T., Al-Khalifa, H.: ”Overview of OSACT4 Arabic offensive language detection shared task”; In Proceedings of the 4th Workshop on open-source arabic corpora and processing tools, with a shared task on offensive language detection, 48-52, (May 2020).

[Mulki et al. 2019] Mulki, H., Haddad, H., Ali, C. B., Alshabani, H.: ”L-hsab: A levantine twitter dataset for hate speech and abusive language”; In Proceedings of the third workshop on abusive language online, 111-118, (Aug 2019).

[Pedregosa et al. 2011] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer,P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher,M., Perrot, M., Duchesnay, É.: ”Scikit-learn: Machine learning in Python”; the Journal of machine Learning research, 12, 2825-2830, (2011).

[Safaya et al. 2020] Safaya, A., Abdullatif, M., Yuret, D.: ”KUISAIL at SemEval-2020 Task 12: BERT-CNN for Offensive Speech Identification in Social Media”; In Proceedings of the Fourteenth Workshop on Semantic Evaluation, (Dec 2020), 2054-2059.

[Shapiro et al. 2022] Shapiro, A., Khalafallah, A., Torki, M.: ”Alexu-aic at arabic hate speech 2022: Contrast to classify”; arXiv preprint arXiv:2207.08557, (2022).

[Soliman et al. 2017] Soliman, A. B., Eissa, K., El-Beltagy, S. R.: ”Aravec: A set of arabic word embedding models for use in arabic nlp”; Procedia Computer Science, 117, 256-265, (2017).

[Talafha et al. 2020] Talafha, B., Ali, M., Za’ter, M. E., Seelawi, H., Tuffaha, I., Samir, M., Farhan,W., Al-Natsheh, H. T.: ”Multi-dialect arabic bert for country-level dialect identification”; arXiv preprint arXiv:2007.05612., (2020).

Journal of Universal Computer Science, vol. 29, no. 11 (2023), 1336-1360
submitted: 19/5/2023, accepted: 30/9/2023, appeared: 28/11/2023 CC BY-ND 4.0

Using Video Activity Reports to Support Remote Project-Based Learning

Kosuke Sasaki
(University of Tsukuba, Ibaraki, Japan
https://orcid.org/0000-0002-1011-8884, ksasaki@slis.tsukuba.ac.jp)

Zhen He
(University of Tsukuba, Ibaraki, Japan, zhe0745@gmail.com)

Tomoo Inoue
(University of Tsukuba, Ibaraki, Japan
https://orcid.org/0000-0003-3600-214X, inoue@slis.tsukuba.ac.jp)

Abstract: Distance learning has been expanding. Learner engagement is particularly important in project-based learning (PBL), but the interaction between teacher and learner and the understanding of learner status, including engagement, is not easy. This study aims to support teacher-learner communication based on learner engagement for remote PBL. In this paper, we propose the use of video activity reports by learners to estimate and understand learner engagement and to demonstrate its feasibility on the basis of the relationship between verbal and nonverbal information that can be obtained from video activity reports and learner engagement. Analysis of 232 video activity reports submitted by eight graduate students while working on remote research-based PBLs reveals that learner engagement decreases (1) when the report contained negative words, (2) when filled pauses were frequent or long, and (3) when silent pauses were infrequent or short. Furthermore, the feasibility of an AI-based support system is demonstrated through the design and implementation of a prototype.

Keywords: Distance learning, Project-based learning, Engagement, Video activity report, Artificial intelligence
Categories: H.5.3, L.3.5, L.3.6
DOI: 10.3897/jucs.113266

1 Introduction

Remote work has grown in popularity in recent years. This involves working without space or time limits [Olson, 1983, Beckel and Fisher, 2022]. It can reduce stress, improve quality of life, and boost well-being [Bailey and Kurland, 2002, Gilson et al., 2015]. Remote work will likely remain popular in the future.

Work engagement is often a focus in the context of work. Work engagement indicates the positive mental state of workers toward their jobs [Kahn, 1990, Schaufeli et al., 2002b, Harter et al., 2002]. It has been shown that when work engagement declines, employees' belongingness to the company, task performance, and motivation are reduced [Harter et al., 2013, Schaufeli and Bakker, 2004]. Managers can help employees maintain their performance when work engagement declines. Therefore, measuring work engagement in remote work is a necessary aspect of managing employees.

Several challenges arise in remote work, including reduced communication opportunities [Mulki et al., 2009, Bezerra et al., 2020, Johnson et al., 2021]. This can affect relationships or task performance and cause mental health issues [De Vincenzi et al., 2022]. It has also been pointed out that lack of communication opportunities can lead to lower work engagement [Galanti et al., 2021].

To address the lack of communication opportunities in remote work, previous studies have focused on activity reports, which are exchanged between managers and employees to report daily operations [Pogorilich, 1992, Lu et al., 2018]. Video activity reports contain verbal and nonverbal information and provide more information than text-based reports. He et al. showed the relationship between video activity reports and engagement, demonstrating the usefulness of video activity reports for remote work [He et al., 2022].

Distance learning is also a form of remote work, as it is a remote activity that is not restricted by time or space. In distance learning, students can learn from home without attending school [Chen et al., 2021] and watch on-demand class videos whenever they wish [Cojocariu et al., 2014, Adnan and Anwar, 2020]. Distance learning has also been criticized as a possible cause of a decline in learner engagement due to reduced communication opportunities [Longhurst et al., 2020, Wilcha et al., 2020, Lee et al., 2020]. Communication support in distance learning is necessary because reduced engagement can lead to negative effects, such as reduced task performance [Tanaka et al., 2021, Zhuang and Chen, 2020].

This study aims to support communication between teachers and learners in distance learning. For our purpose, we propose using video activity reports to help teachers estimate and understand learner engagement. It is expected that teachers will be able to increase communication opportunities between themselves and students whose engagement is declining by understanding learner engagement. Increased communication opportunities are also expected to lead to increased learner engagement. This study focuses on the learner engagement of students engaged in project-based learning (PBL).

PBL is a pedagogical approach that combines knowledge acquisition and problem solving [Markham, 2011]. To demonstrate the effectiveness of PBL in computer science, McManus and Costello conducted PBL to ask students to investigate the feasibility of an unmanned aerial vehicle (UAV) at low monetary cost and to implement it. As a result, students acquired not only specialized knowledge and skills but also time management and presentation skills [McManus and Costello, 2019]. In addition, Chounta et al. employed PBL in a camping-style setting, in which small teams worked on programming tasks. Every team that participated in this PBL received high assessments in terms of creativity and novelty, among other criteria [Chounta et al., 2017]. A common point in these examples of PBL is that learners progressed their learning activities on their own. PBL requires learners to not only complete tasks given by the teacher but also find their own tasks and perform their own learning activities with the support of teachers [Markham, 2011]. Therefore, learner engagement, which indicates a positive mental state toward a task, is essential for the success of PBL.

Similarly, in the present study, students determined a research theme, learned research methods, and proceeded with a research project while applying the methods. Although the teacher was a supervisor, they did not direct all of the students' tasks. Thus, the activity in this study can be regarded as PBL. This study focused on learner engagement over several months during the PBL period.

To understand the learner engagement of students engaged in distance PBL using video activity reports, it is necessary to clarify the relationship between video activity reports and learner engagement. In this study, we analyzed the relationship between 232 video activity reports of eight graduate students obtained by He et al. [He et al., 2022] and learner engagement on that day.

We also created a system prototype to help teachers. We discussed how the findings of this study can be utilized and what aspects should be the focus of the system.

The contributions of this study are as follows.

- Based on the verbal and nonverbal information in the video activity report, we revealed the following relationship regarding the engagement of students engaged in PBL activities:
  - Engagement declines when negative phrases appeared in video activity reports.
  - Engagement declines when the frequency of filled pauses is high or when the duration of filled pauses is long.
  - Engagement declines when the frequency of silent pauses is low or when the duration of silent pauses is short.
- Demonstrated a system prototype that allows teachers to grasp learner engagement using video activity report and showed prospects of the system using AI technology.

2 Related Work

2.1 Issues in Remote Work

Studies comparing work in face-to-face environments with remote work have been conducted in recent years. It has been reported that the diversity of home environments and life rhythms in teleworking must be considered to avoid failure in work coordination and communication [Breideband et al., 2022]. Other studies have claimed that remote activities between teams do not always work well, even if they work well within a team [Hu et al., 2022], and that office work and work-from-home work cannot be coordinated in the same way [de Souza Santos and Ralph, 2022]. These studies have pointed out that communication and coordination issues related to remote work.

One problem with remote work is the lack of communication opportunities [Mulki et al., 2009, Bezerra et al., 2020, Johnson et al., 2021]. Communication between teachers and students in remote environments may also be reduced, which impedes learning activities. When collaborating with multiple people, teamwork and collaboration can be inhibited in a remote environment [Sarker et al., 2012, Pyöriä, 2011, Baruch, 2000, Pearlson and Saunders, 2001]. It has also been suggested that poor communication may hinder team morale [Overmyer, 2011, Tremblay and Thomsin, 2012] or reduce team commitment [Golden, 2009].

These studies indicate that it is necessary to support communication in distance learning.

2.2 Engagement

One measure of mental state is engagement, which relates to motivation [Kahn, 1990, Schaufeli et al., 2002b, Harter et al., 2002]. Motivation is the willingness to perform a task, whereas engagement indicates positivity when immersed in a task [Afflerbach and Harrison, 2017]. High engagement indicates a positive mental state, leading to better task performance [Harter et al., 2013] and increased organizational belonging [Schaufeli and Bakker, 2004].

Many studies in the field of education have focused on engagement as a noncognitive factor that influences learning [McGill et al., 2019]. In the educational context, this has commonly been referred to as "student engagement" since the 1980s [Trowler, 2010]. Marks defined it as "the attention, interest, investment, and effort students expend in the work of learning" [Marks, 2000]. Learner engagement contributes to social and cognitive development and improves academic performance [Finn and Zimmer, 2012, Schaufeli et al., 2002a, Loscalzo and Giannini, 2019, Dimitriadou et al., 2020], while reduced engagement increases dropout risk [Schaufeli et al., 2002a, Rumberger, 1987]. These studies show that learner engagement is worthy of attention in the field of education.

There are different definitions of engagement in different fields, including education and psychology [Azevedo, 2015]. This study follows the definition by Schaufeli et al., who state that "engagement is characterized by vigor (high activation) and dedication (high identification). Furthermore, ..., and engagement includes absorption" [Schaufeli et al., 2002b]. This definition is often used in computer-supported cooperative work. They proposed that engagement is the opposite of burnout. Vigor (high activation) and dedication (high identification), which are the components of engagement, are the opposite of exhaustion (low activation) and cynicism (low identification), which are two of the three components of burnout. Absorption, the third component of engagement, which was added based on the interview survey results, corresponds to reduced efficacy, the third component of burnout [Schaufeli et al., 2002b].

Engagement is affected by job resources (e.g., autonomy, feedback, and social support), personal resources (e.g., optimism, self-efficacy, and resilience), and job demands (e.g., work pressure and physical/mental demands). Demerouti et al. proposed the JD-R model, which suggests that an imbalance between the two resources and demands can cause burnout [Demerouti et al., 2001]. Bakker and Demerouti claimed that engagement is high when job and personal resources are adequate and job demands are in balanced [Bakker and Demerouti, 2008].

Engagement as originally proposed by Schaufeli et al. was used in the work context [Schaufeli et al., 2002b]. However, Schaufeli et al. showed that their proposed engagement is also meaningful for learners. In particular, they showed that higher vigor, a component of engagement, can lead to higher academic performance [Schaufeli et al., 2002a].

In this study, we used the definition of engagement proposed by Schaufeli et al. [Schaufeli et al., 2002a] to understand the engagement of learners engaged in distance PBL.

2.3 Measuring Engagement

Since engagement is an abstract concept, questionnaires were used to measure it. To measure engagement, Schaufeli et al. created the Utrecht Work Engagement Scale (UWES) with 17 questions focusing on vigor, deduction, and absorption [Schaufeli et al., 2002b]. UWES-S for students [Schaufeli et al., 2002a], UWES-9 with nine questions [Schaufeli et al., 2006], and UWES-3 with three questions [Schaufeli et al., 2017] have also been proposed.

Measuring engagement changes is challenging, as engagement changes daily [Breevaart et al., 2012, Sonnentag et al., 2010]. Engagement surveys are often conducted annually, making it difficult to track changes [Shami et al., 2015]. A decrease in engagement may decrease engagement in others [Mitra et al., 2017]. Thus, methods are required to measure engagement over short periods so that teachers can quickly detect declining learner engagement.

However, frequent surveys can burden respondents, necessitating alternative engagement measurement methods based on daily activities. For example, Kajiwara et al. measured work engagement using pulse and eye and body movements in a multimodal environment [Kajiwara et al., 2019]. To measure engagement in this study, we focused on daily activity reports in remote PBL.

2.4 Activity Report

Activity reports are used by teachers to track student progress or by students to consult with teachers. Chickering and Gamson argued that student-teacher contact and prompt feedback can improve undergraduate education [Chickering and Gamson, 1987]. Etkina and Harper's study of undergraduate education found that weekly reports help students solve problems quickly and reflect on their learning and help teachers adjust their teaching methods [Etkina and Harper, 2002]. Ito et al. found that weekly reports could be used to track student learning [Ito et al., 2016].

Activity reports are generally text based. Activity reports in remote work can be sent via text-based platforms such as email, Slack[1], Microsoft Teams[2], and other instant messaging apps [Lu et al., 2018]. Systems that use emojis to express emotions [Tang et al., 2022] and AI to present emotion analysis of text chats to participants [Nguyen et al., 2022] have been developed. Some studies have focused on enhancing communication by introducing chatbots [Benke et al., 2020a] and exploring the characteristics of effective remote teams based on text communication analysis [Cao et al., 2021]. These studies suggest that text-based activity reports support remote work.

Other studies used text activity reports to measure engagement. Tanaka et al. developed a learning model to estimate engagement based on text chat frequency and worker affiliation [Tanaka et al., 2021]. Shami et al. and Golestani et al. demonstrated that internal SNSs content can measure engagement [Shami et al., 2015, Golestani et al., 2018]. Cao et al. found that more exclusive words and second-person pronouns in text chats indicated better online team performance [Cao et al., 2021]. Wang et al. proposed a method for estimating user performance in Slack channels by analyzing messages and using content-independent features [Wang et al., 2022].

However, textual communication can induce negative emotions in users owing to a lack of nonverbal information [Hassib et al., 2017, Benke et al., 2020b]. Thus, we considered audio and video media. While voice can convey messages faster than text [El-Shinnawy and Markus, 1997], and can reduce the decision-making time when used in meetings accordingly [Suh, 1999], video is better for conveying intentions and emotions based on users' facial expressions, gestures, and facial movements [Veinott et al., 1999, Bänziger et al., 2009]. Because this study focuses on engagement, which indicates a mental state, we used video activity reports, which are more likely to convey mental information such as emotions.

2.5 Video Activity Report

He et al. studied an employee's engagement using video activity reports. Over two years, 418 reports were collected and analyzed, focusing on the employee's filled and silent pauses in the reports. The results showed that when engagement was high, filled pauses

[1] Where work happens | Slack, https://slack.com/ (Visited on January 31, 2023)

[2] Video Conferencing, Meetings, Calling | Microsoft Teams, https://teams.com/ (Visited on January 31, 2023)

were low; when satisfaction was high, filled pauses were short; when clarity was high, filled pauses were minor; and when job ratings were high, filled pauses were low and silent pauses were longer [He et al., 2020a]. This suggests that video activity reports, which include nonverbal information, may provide insights into engagement and other mental states.

He et al. studied video activity reports of students engaging in PBL, obtaining 232 reports from eight students over four months and interviewing six students. The learner comments suggested the availability of video activity reports in PBL [He et al., 2022]. He's analysis of nonverbal information in the reports showed that filled and silent pauses were related to engagement [He, 2022].

2.6 Filled and Silent Pauses

Filled and silent pauses are paralinguistic cues that express emotions and intentions [Fujisaki, 2004]. They are represented by pitch, intensity, and speech ratio, among others [Johar, 2014]. Some studies examined emotion estimation using paralinguistic cues [Delaborde and Devillers, 2010, Wang and Hu, 2018].

Christenfeld and Creager suggested that filled pauses occur when a speaker's thought process cannot keep up with their speech process [Christenfeld and Creager, 1996]. Goto et al. claimed that speakers employ filled pauses to make time for the next utterance to develop and maintain the conversation and that filled pauses express unconscious mental or thought states, such as self-confidence, hesitancy, anxiety, or modesty [Goto et al., 1999]. Filled and silent pauses are influenced by a speaker's levels of confidence and anxiety [Pope et al., 1970]. When the speech content is more abstract, speakers have more filled and silent pauses [Rochester, 1973]. When the speech content is more complex, the speaker needs longer time to start speaking [Brennan and Williams, 1995]. Moreover, filled and silent pauses are also affected by the situation. Filled pauses are common in interviews but not in speeches [Duez, 1982, Duez, 1985]. More silent pauses occur in stressful environments than in non-stressful ones [Buchanan et al., 2014].

These studies have shown that filled and silent pauses reveal a speaker's mental state. We explored if the pauses can be used to understand learner engagement.

3 Method

3.1 Research Questions

This study aims to support communication between teachers and students in distance PBL based on learner engagement. Learner engagement is related to task performance and dropout risk. Therefore, it is preferable for teachers to focus on learner engagement. However, it is difficult to measure engagement remotely due to the fewer communication opportunities than in face-to-face environments. A system is required to help teachers understand learner engagement in distance learning.

This study constructs a mechanism for teachers to understand learner engagement in distant environments. It is necessary to identify the factors that teachers should focus on in order to understand learner engagement. Therefore, this study focuses on video activity reports to understand learner engagement. An activity report is a teacher-student communication tool used in distance learning, which is usually sent via text. Video activity reports contain nonverbal information related to learner mental state and engagement.

We analyzed the video activity reports obtained by He et al. [He et al., 2022] of students engaged in PBL in a distance environment to investigate the relationship between the reports and engagement. Focusing on two types of information obtained from video activity reports, that is, verbal and nonverbal, we set the following research questions:

RQ1: How does learner engagement relate to verbal information within video activity reports?

RQ2: How does learner engagement relate to nonverbal information within video activity reports?

RQ1 focuses on the relationship between verbal information and engagement in video activity reports. Previous studies on video activity reports have not focused on report content [He et al., 2020a, He et al., 2022, He, 2022]; thus, it is essential to investigate how report content impacts engagement.

Engagement decreases in negative mental states. A person using negative statements has negative emotions [Kahn et al., 2007]. Therefore, we focused on negative words or phrases in video activity reports. We formulated the following hypothesis for RQ1:

H1: Negative statements in video activity reports indicate declining learner engagement.

RQ2 focuses on the relationship between nonverbal information and engagement in video activity reports. He et al. studied filled and silent pauses in employee reports. Filled pauses are related to the speaker's mental state [Goto et al., 1999], and silent pauses can be caused by mental stress [Lee et al., 2017]. Both pauses may be related to users' engagement.

The previous study analyzed the video activity reports of only one participant, so its conclusions are not generalizable [He et al., 2020a]. Our study analyzed video activity reports from eight students to examine the relationship between filled/silent pauses and engagement. We formulated the following hypotheses for RQ2:

H2-1: Occurrence of filled pauses in video activity reports relates to engagement.

H2-2: Occurrence of silent pauses in video activity reports relates to engagement.

3.2 Video Activity Report and Engagement Score

This section describes the collection of video activity reports [He et al., 2022].

3.2.1 Participants

The Ethics Committee on Library, Information and Media Studies at University of Tsukuba approved the previous study (No. 20-32) [He et al., 2022]. The experiment collected video activity reports and used snowball sampling on SNS (WeChat[3] and LINE[4]) to recruit eight male graduate students aged 23–26, as shown in Table 1.

[3] WeChat - Free messaging and calling app,https://www.wechat.com/ (Visited on January 31, 2023)

[4] LINE ｜ always at your side., https://line.me/en/ (Visited on January 31, 2023)

Participant	Age	Gender	Nationality	# of submitted reports / Actual trial days
P1	23	Male	Chinese	56 / 63
P2	25		Chinese	23 / 24
P3	25		Chinese	33 / 36
P4	25		Chinese	33 / 34
P5	23		Chinese	37 / 37
P6	24		Chinese	16 / 16
P7	26		Chinese	16 / 16
P8	23		Japanese	18 / 26

Table 1: Participants of the study

No.	Factor	Content
1	Vigor	When I'm doing my work as a student, I feel like I am bursting with energy.
2	Dedication	I am enthusiastic about my studies.
3	Absorption	I am immersed in my studies.

Table 2: Questionnaire items

We verified that, at minimum, non-native Japanese speakers could communicate in Japanese at the N2 level of the Japanese-Language Proficiency Test[5] and could easily report their activities in Japanese.

All participants engaged in PBL remotely, similar to teleworking, during their participation period. In this experiment, participants were asked to submit a video activity report of their activities and answer a questionnaire to measure their engagement every weekday. Only when participants submitted a video activity report and a response to the questionnaire were they paid a gratuity of 100 yen per day. Table 1 presents the number of times the participants submitted the data during their participation in the experiment. The experiment was conducted between August 2020 and November 2020.

3.2.2 Engagement Score

To measure engagement, we used the UWES-3 questionnaire proposed by Schaufeli et al. [Schaufeli et al., 2017], as it has the fewest questions, making daily surveys less burdensome. Because the UWES-3 was designed for company workers, the wording of some of the UWES-3 questions was changed by referring to the UWES-S questionnaire, an engagement measurement questionnaire for students [Schaufeli et al., 2002a]. Table 2 shows the actual questions.

The participants answered each question on a scale of 0 to 6 (0=never, 6=always). The average of the three questions was defined as the "engagement score" for the day, which we used in our analysis.

[5] JLPT Japanese-Language Proficiency Test, https://www.jlpt.jp/e/) (Visited on January 20, 2023)

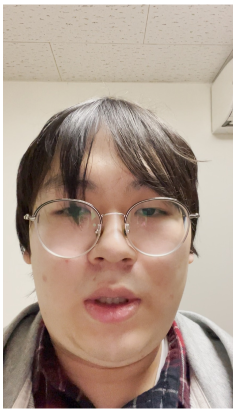


Figure 1: A screenshot of a submitted video report

3.2.3 Procedure

During the experiment, the participants were asked to submit a video activity report and answer the questionnaire on weekdays, excluding holidays. They were given the following instructions:

- Your video activity report should be taken with your laptop, smartphone, or any other device with a microphone and camera.
- Any objects that mask your face should not be worn, and the video should be taken in a bust shot. (Figure 1 shows a screenshot of the actual video submitted)
- The report's content should be related to your study or research, such as what you learned in the day's lectures, how they prepared your reports, and the progress of your research.
- There is no limit to the length of each video. Each video will be approximately 30 seconds long, but it is acceptable for the video to be shorter or longer than 30 seconds.
- There are no restrictions on where you take video, but try to record in a place that is as quiet as possible. In addition, ensure that your voice is clear.

This procedure was based on the previous study in which similar video activity reports were collected [He et al., 2020a].

We asked participants to record a video activity report per day for two reasons. First, other studies that focused on activity reports also collected daily reports (e.g., [Pogorilich, 1992, Lu et al., 2018]). Second, engagement may vary from day to day [Breevaart et al., 2012, Sonnentag et al., 2010].

We set a guideline of 30 seconds as the length of each video activity report based on a trial conducted by the experimenter without recording time limitation in advance, which resulted in an average of approximately 30 seconds.

Participants were asked to take a video activity report and answer the questionnaire after recording the report every weekday. They uploaded the recorded videos once a

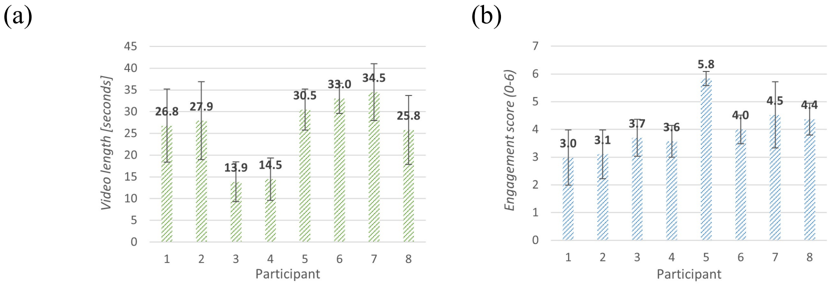


Figure 2: (a) Average length of video reports / (b) Engagement score. Error bars on each graph show the standard errors.

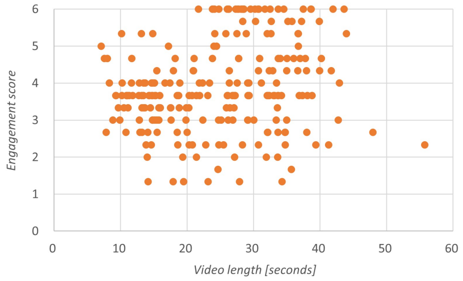


Figure 3: Relationship between engagement score and length of video activity reports

month to cloud storage. The engagement score for the day was calculated based on the answers to the questionnaire and was collected with the same day's video activity report.

In addition, participants were asked to continue participating in the experiment for at least four weeks when they consented to the experiment. Four weeks after the start of the experiment, participants could withdraw from the experiment at any time.

4 Result

Table 1 lists the 232 video activity reports and questionnaire responses. Video activity reports and engagement scores were always collected together. We investigated the relationship between the video activity report and the engagement score on the same day. Participant participation ranged from 16 to 63 weekdays. Figure 2 shows the video activity report length and engagement score for each participant. The average video activity report was 24.8 seconds (SD: 9.6 seconds), with an average and median engagement score of 3.9 and 3.7 (SD: 1.2), respectively.

Engagement scores below the median were labelled "low engagement score," and those above the median were labelled "high engagement score." Of the 232 scores, 128 were classified as low engagement scores and 104 as high engagement scores.

4.1 Length of Video Activity Reports and Engagement Score

We investigated the relationship between the video length and engagement scores. Figure 3 shows a scatter plot of video length and engagement score. Pearson's correlation

My task is to create a system, but that is still **not** finished.
We tried to solve the problem we had last week, but we have **not** solved it yet.
My research is underway but **not** progressing very well.
I **cannot** give you concrete results.
I am not feeling well today, so I went to the hospital, and I **did not** perform any tasks.
I have looked for papers related to my research but have **not** found helpful ones.
The model was trained in an outdoor environment; therefore, it was found to be **less** accurate in an indoor environment.
I read the paper, but I am **not** very good at English, so I had to translate it and use online sites to complete it.
This analysis software is a bit **difficult** to use, and I am learning and analyzing various functions.
I tried to analyze the data, but still had **insufficient** data to analyze.

Table 3: Examples of negative expressions in video activity reports. ***Bolded phrases*** *indicate the basis for determining that the report was negative.*

coefficient was not high but correlated ($r = .22, t(230) = 3.46, p = .0006$).

Longer video activity reports were more possibly related to higher engagement scores.

4.2 Negative Statements and Engagement

We analyzed the engagement scores and content of the video activity report (RQ1).

We used Google Cloud Speech-to-Text API[6] to transcribe the participants' video reports into Japanese. All errors were manually corrected.

The labelers determined whether the video activity report included negative expressions or not. Table 3 lists several negative phrases that were found in the reports. A video activity report was labeled = negative if (1) a negating word for the verb of the sentence in which the student was the subject appeared once or (2) a generally negative or pessimistic term appeared. If no negative words appeared, the report was labeled as non-negative. In cases in which the judges of the two labelers differed, labeling was performed by consensus. Of the 232 video activity reports, 41 were labeled as negative and 191 as non-negative. The agreement rate was 99.6% with labelers only disagreeing on one case. We think the reason for this very high agreement rate was that enough detailed criteria were predetermined to avoid subjective differences in labeling.

Figure 4 shows the engagement scores when video activity reports labeled negative or non-negative were submitted. The average engagement score for reports labeled negative was 3.4 (SD: 1.2), and 4.0 (SD: 1.2) for reports labeled non-negative. Welch's t-test ($t(58) = -2.78, p = .007 < .01$ $(Cohen's\ d = 0.48)$) revealed a significant difference, suggesting that when a video activity report containing negative phrases was submitted, engagement on that day was lower than when a report containing no negative words was submitted.

This result supports hypothesis H1, "Negative statements in video activity reports indicate declining learner engagement."

[6] Speech-to-Text: Automatic Speech Recognition | Google Cloud, https://cloud.google.com/speech-to-text (Visited on January 20, 2023)

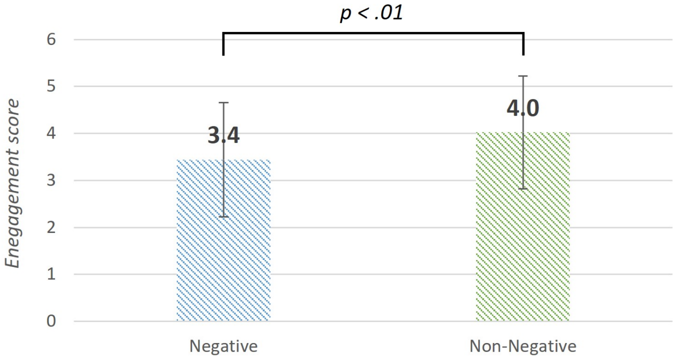


Figure 4: Average of engagement score on days when negative or non-negative video activity reports were submitted

4.3 Nonverbal Information and Engagement

The following analysis was conducted to clarify the relationship between learner engagement and nonverbal information in video activity reports (RQ2).

The experimenters manually labelled the video activity reports for filled and silent pauses. Based on previous studies [Goto et al., 1999, Campione and Véronis, 2002], typical Japanese fillers, such as /ee-/, /maa-/, and /ano-/, as well as those with the end of the word extended (e.g., /kyo wa-/ ("Today...")), were labelled as filled pauses. Silent pauses lasting for more than one second after the end of the utterance were labelled. ELAN[7] was used for labelling.

Table 4 lists the items analyzed for nonverbal information. The frequency, average length per minute, and length per time were calculated for filled and silent pauses.

Figure 5 shows the frequency and time per minute of filled pauses and time per filled pause for the days with low and high engagement scores (N=232). We compared whether there was a difference in filled pauses between the days with low and high engagement scores.

The frequency of filled pauses (Figure 5(a)) averaged 11.3 times/minute (SD: 6.1 times/minute) for the days with low engagement scores and 9.3 times/minute (SD: 7.7 times/minute) for the days with high engagement scores. Welch's t-test showed that $t(194) = 2.17, p = .03 < .05$ $(Cohen's\ d = 0.29)$, indicating a significant difference.

The average length of filled pauses per minute (Figure 5(b)) was 5.0 seconds (SD: 3.4 seconds) on average for the days with low engagement scores and 3.6 seconds (SD: 3.3 seconds) for the days with high engagement scores. Welch's t-test showed that $t(223) = 3.29, p = .001 < .01$ $(Cohen's\ d = 0.43)$, indicating a significant difference.

Furthermore, the average length of filled pause per time (Figure 5(c)) was 0.5 seconds (SD: 0.3 seconds) for the days with low engagement scores and 0.4 seconds (SD: 0.3 seconds) for the days with high engagement scores. Welch's t-test showed that $t(931) = 4.24, p < .01$ $(Cohen's\ d = 0.27)$, indicating a significant difference.

These results support hypothesis H2-1, "Occurrence of filled pauses in video activity report relates to engagement." Specifically, the number or duration of filled pauses per

[7] ELAN | The Language Archive, https://archive.mpi.nl/tla/elan (Visited on January 31, 2023)

Item	Definition	Unit	Average (SD)
Frequency of filled pauses	Number of filled pauses per minute in each video activity report	times/minute	10.4 (6.9)
Length of filled pauses	Length of filled pauses per minute in each video activity report	second	5.0 (3.2)
Length of each filled pause	Length of filled pauses per time	second	0.4 (0.3)
Frequency of silent pauses	Number of silent pauses per minute in each video activity report	times/minute	2.6 (3.5)
Length of silent pauses	Length of silent pauses per minute in each video activity report	second	8.3 (4.7)
Length of each silent pause	Length of silent pause per time	second	1.5 (0.5)

Table 4: Measured non-verbal information

minute or the duration per time was greater in video activity reports on days with low engagement scores than on days with high engagement scores.

Figure 6 shows the frequency and time per minute of silent pauses and time per silent pause for the days with low and high engagement scores (N=232). We examined whether there was a difference in the occurrence of silent pauses between days with low and high engagement scores.

The frequency of silent pauses (Figure 6(a)) averaged 1.6 times/minute (SD: 3.0 times/minute) for the days with low engagement scores and 3.8 times/minute (SD: 3.8 times/minute) for the days with high engagement scores. Welch's t-test showed that $t(192) = -4.82, p < .01$ $(Cohen's\ d = 0.65)$, indicating a significant difference.

The average length of silent pauses per minute (Figure 6(b)) was 2.4 seconds (SD: 4.4 seconds) for the days with low engagement scores and 5.8 seconds (SD: 5.6 seconds) for the days with high engagement scores. Welch's t-test showed that $t(191) = -5.16, p < .01$ $(Cohen's\ d = 0.69)$, indicating a significant difference.

Furthermore, the average length of silent pauses per time (Figure 6(c)) was 1.4 seconds (SD: 0.5 seconds) for the days with low engagement scores and 1.6 seconds (SD: 0.5 seconds) for the days with high engagement scores. The Welch's t-test of $t(166) = -1.62, p = .11$ $(Cohen's\ d = 0.21)$ showed no significant difference.

Thus, hypothesis H2-2, "Occurrence of silent pauses in video activity report relates to engagement" is supported. Specifically, reports with low engagement scores had fewer or shorter silent pauses per minute than those with high engagement scores did.

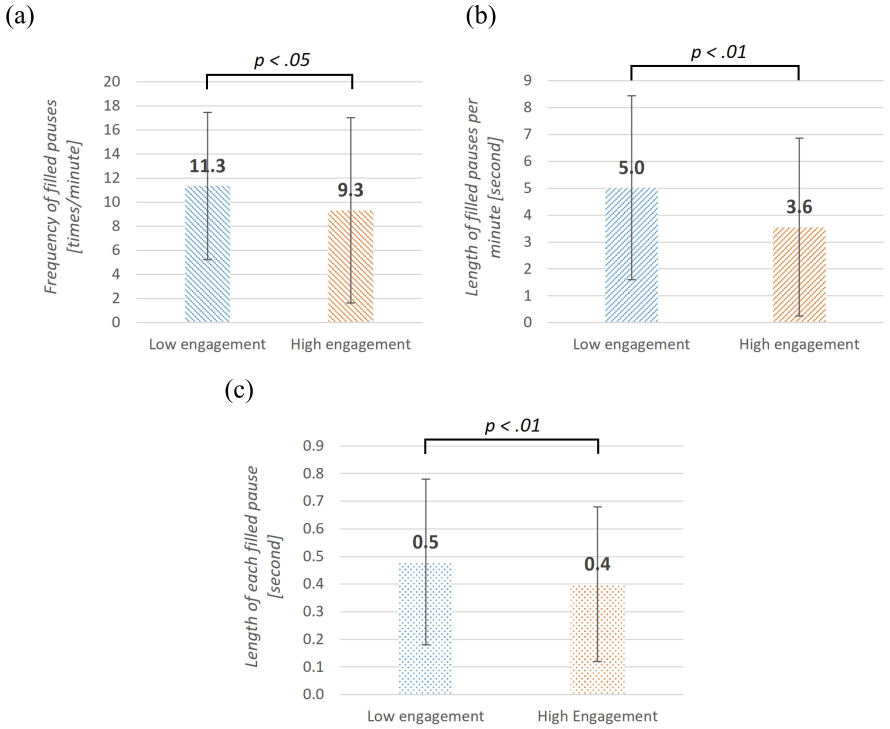


Figure 5: (a) Frequency of filled pauses of two groups based on engagement score. (b) Length of filled pauses of the two groups. (c) Length of each filled pause of the two groups

5 Discussion

This study will create a framework to help teacher-student communication by understanding learner engagement in distance learning. This chapter summarizes the findings, introduces a system prototype, and describes the study's limitations and future research.

5.1 Findings

Based on the results, the answers to the research questions are summarized as follows. The answer to RQ1, "How does learner engagement relate to verbal information within video activity reports?" is that when negative phrases were included in the video activity report, learner engagement was lower than when negative phrases were not included in the video activity report.

The answer to RQ2, "How does learner engagement relate to nonverbal information within video activity reports?" is that video activity reports on days with low engagement scores had a greater number or longer duration of filled pauses or a smaller number or shorter duration of silent pauses per minute than those on days with high engagement scores.

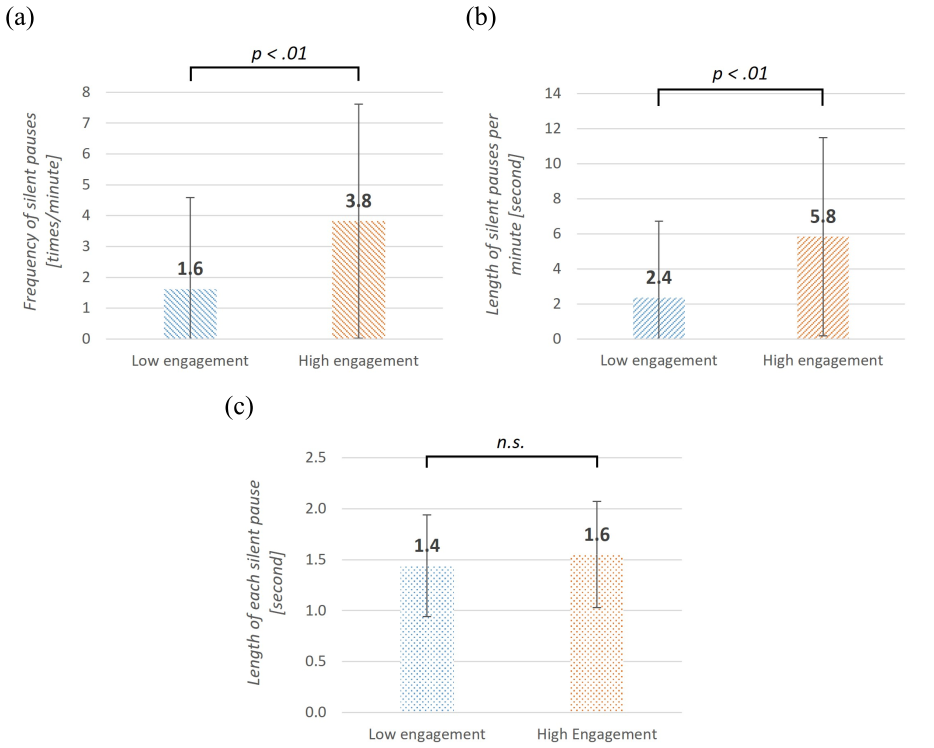


Figure 6: (a) Frequency of silent pauses of two groups based on engagement score. (b) Length of silent pauses per minute of the two groups. (c) Length of each silent pause of the two groups

These findings are useful for building a system that uses video activity reports to understand learner engagement. For instance, if the system identifies negative phrases or frequent filled pauses, it indicates a decline in engagement.

Filled and silent pauses are affected by cognitive, mental, and speech situation, and it is difficult to determine whether these factors are positively or negatively correlated. We observed graduate students engaged in PBL, which is based on research activities, and recorded their reports. We discovered that when learner engagement was low, filled pauses were more frequent or longer and silent pauses were infrequent or shorter. The reasons for these are not yet known, but a possible explanation might be that during the video activity report, students employ filled pauses when thinking of what to say and use silent pauses when nothing to say [He et al., 2022].

A small positive correlation was found between video activity report length and engagement score. It is possible that activity reports become longer due to an increase in the number of items students want to report as the learning activity progresses, which could indicate high engagement. Additionally, whether and how much the length guideline of 30 seconds that was given to the participants affected engagement can be examined in the future.

5.2 Classification of Activity Reports Toward Automated Engagement Estimation

We labeled a report as negative if one negative term was found. This judgement may be easy to implement for a support system. Consider the following report obtained in this study, for example.

> *Today, I continued a survey on OpenFace [an analysis toolkit]. I sought to determine how the data from OpenFace could be used. We received a lot of data from OpenFace, so I will continue to investigate this because I* ***do not yet*** *know which parts of the data can be effectively used in my research. That is all.* (P1)

This report was labeled as negative according to the phrase "**do not yet**". However, this does not necessarily mean the report itself has a negative tone.

As a step toward automated engagement estimation that would work in an support system, we also conducted a preliminary automated sentiment analysis of the reports to investigate current applicability of the analysis method. We implemented a standard sentiment analysis program using BERT [Devlin et al., 2019], a natural language model widely used today for natural language processing; cl-tohoku/bert-base-japanesewhole-word-masking[8], a standard pre-trained Japanese BERT model to tokenize words to input into BERT; and a standard fine-tuned Japanese sensentiment analysis model, koheiduck/bert-japanese-finetuned-sentiment[9], which classifies sentence as POSITIVE, NEGATIVE, or NEUTRAL.

Of the 232 activity reports, 128 were classified as POSITIVE, 13 as NEGATIVE, and 91 as NEUTRAL as a result. Considering negative reports, 39 were labeled negative by the labelers but not by automated sentiment analysis, 11 were judged negative by automated analysis but not by the labelers, and two were judged negative by both. Machine learning judgments of the reports differed significantly.

The accuracy of the automated sentiment analysis should be mentioned. Reports judged as NEGATIVE by sentiment analysis include the following:

- *Today, I was working on English slides for my CollabTech [an international conference] presentation. My current progress is that I completed half of the slides on the research background. I plan to complete all of it by the end of this week. That's all.* (P1)
- *Today, I sorted out the questions I received at the workshop and other mentions on the form that were not asked during my presentation.* (P2)
- *Today, I attended the seminar and got advice from the professor on how to survey and on the main points [for my research].* (P4)

These examples are classified as NEGATIVE in the automated analysis, although they do not seem to be negative in their meanings. It is still difficult to analyze sentiments automatically in Japanese activity reports.

[8] https://huggingface.co/cl-tohoku/bert-base-japanese-whole-word-masking (Visited on August 16, 2023)

[9] https://huggingface.co/koheiduck/bert-japanese-finetuned-sentiment (Visited on August 16, 2023)

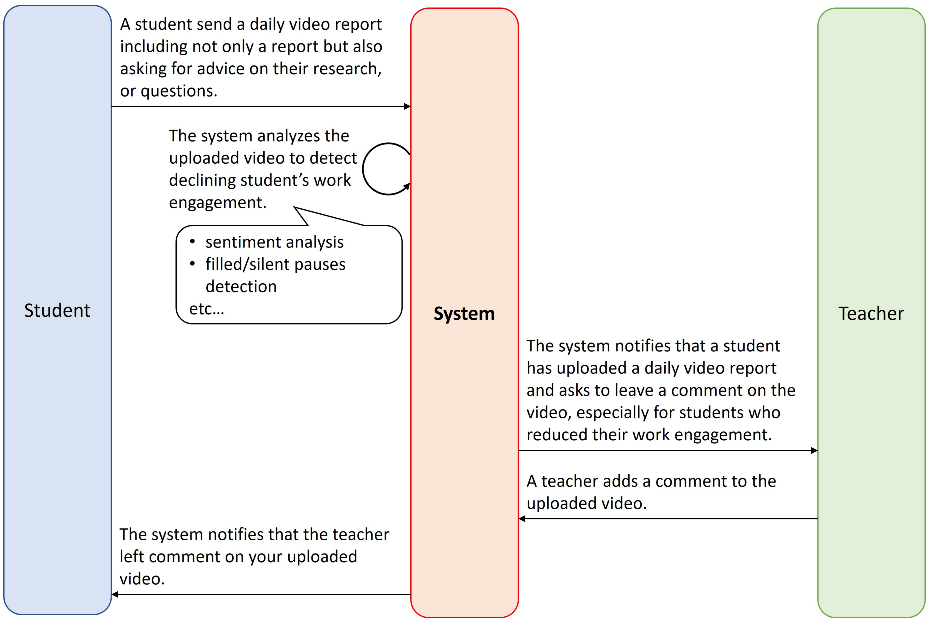

Student
A student send a daily video report including not only a report but also asking for advice on their research, or questions.
The system analyzes the uploaded video to detect declining student's work engagement.
• sentiment analysis
• filled/silent pauses detection
etc...
System
Teacher
The system notifies that a student has uploaded a daily video report and asks to leave a comment on the video, especially for students who reduced their work engagement.
A teacher adds a comment to the uploaded video.
The system notifies that the teacher left comment on your uploaded video.

Figure 7: A communication system between a teacher and a student

5.3 Communication System Estimating of Engagement Using Video Activity Report

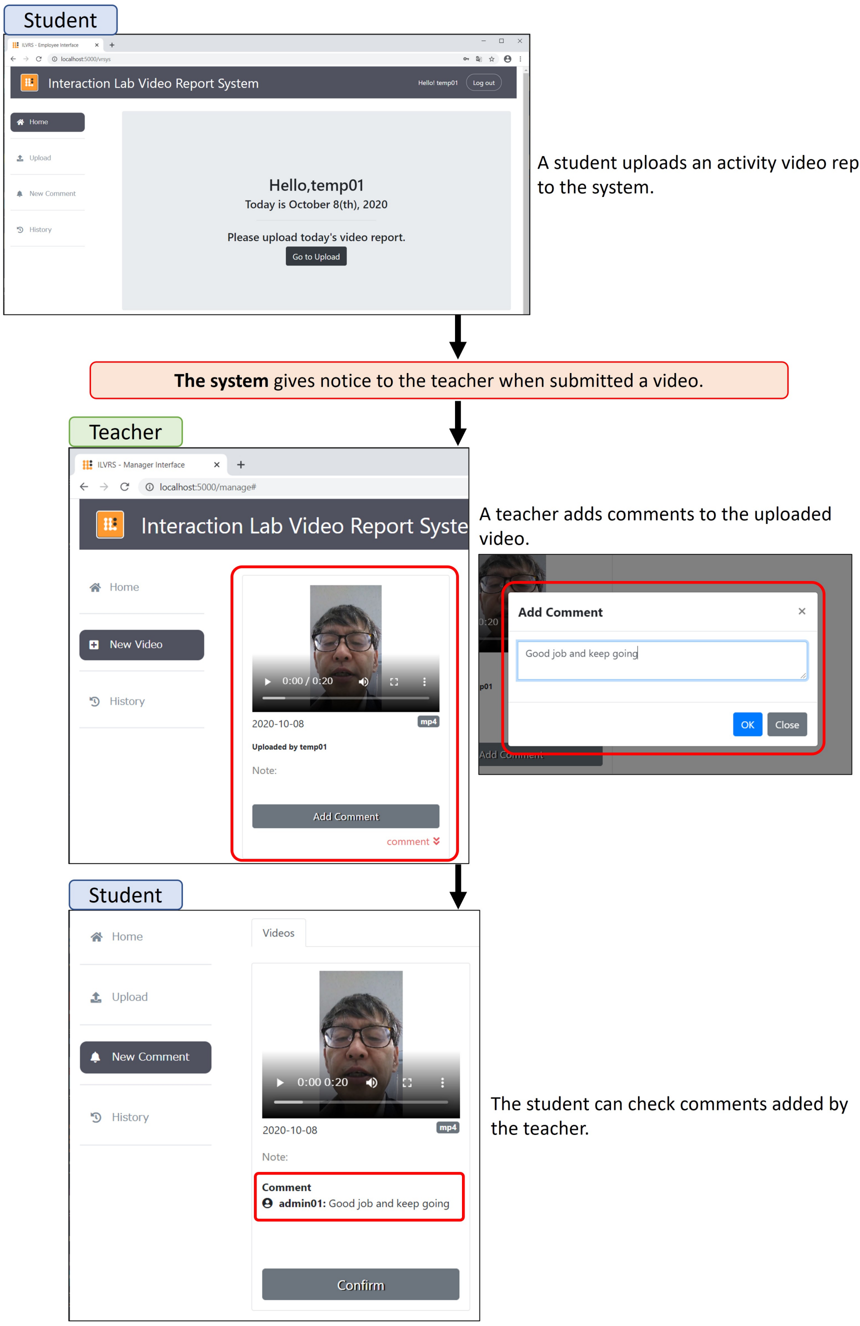
Figure 7 shows a web-based system for teachers to track learner engagement. Students upload video activity reports asking their teachers to comment. The student is notified when the teacher adds comments, allowing them to review the feedback.

A system prototype is shown in Figure 8 [He et al., 2020b]. This system allows remote communication between teachers and students, based on video activity reports. However, this system is unwieldly for teachers. Generally, teachers manage multiple students. Therefore, reviewing video activity reports and commenting on them requires a significant amount of time.

The results of this study can be used in this system. The system can detect students with declining engagement by automatically analyzing video activity reports. Notifying the teachers of such students allows them to focus on those who need more attention while saving time.

5.4 Considering How to Estimate Engagement with Greater Accuracy

This study focused on negative phrases, filled pauses, and silent pauses in order to measure engagement. Other factors, such as part-of-speech or the type of words, can also be considered as verbal information [Wang et al., 2022, Cao et al., 2021]. Nonverbal information can also be considered, such as speech style (e.g., inflection and volume), eye contact, and body movement. In addition, metadata, such as the time, frequency, and duration of report submission, can be focused on.



Hello,temp01

Today is October 8(th), 2020

Please upload today's video report.

A student uploads an activity video report to the system.

The system gives notice to the teacher when submitted a video.

2020-10-08

Uploaded by temp01

Note:

Good job and keep going

A teacher adds comments to the uploaded video.

2020-10-08

Note:

Comment

admin01: Good job and keep going

The student can check comments added by the teacher.

Figure 8: The prototype system

It is possible that estimating engagement with multiple factors could lead to more precise results, but manually determining the useful factors is time-consuming. AI technology, such as deep learning, can identify the features required for a more accurate engagement estimation.

In this study, we found that many filled pauses indicated low engagement but not always for all students. Therefore, to accurately measure engagement, individual user speech and body movements should be observed. In a system used by many students, AI technology can be used to understand individual behavior, which is difficult with human resources. AI can create a model for each user's behavior. When a user's actions deviate from this model, changes, such as decreased engagement, may occur. He et al. observed that students performed unusual movements in video activity reports, such as frequently lifting their head [He et al., 2022]. There is a possibility that unusual movements could indicate engagement. AI can learn a user's typical behavior, such as head movement, eye direction, and voice tone, from video activity reports to detect unusual movements.

In summary, AI technology can estimate learner engagement more accurately. Deep learning based on video activity reports can identify key features, such as negative phrases, filled pauses, or silent pauses. Additionally, a system that creates a model of each student's behavior to detect unusual actions can help estimate engagement based on their habits.

5.5 Limitation and Future Work

In this study, video activity reports obtained from eight students were analyzed. Because these eight students came from the same laboratory of the same major, the results might include biases accordingly. If students from different laboratories, genders, or nationalities participate, the results might be somewhat different.

Several future studies are needed to obtain a high level of engagement accuracy. The first is obtaining individual behavioral changes to reduce students' environmental differences. As Azevedo reported, using self-report measures alone is inadequate for measurement engagement [Azevedo, 2015]. AI technology helps understand individual behavior. For teachers to grasp learner engagement more accurately, it will be necessary to combine what this study has found with what AI reveals by focusing on each individual change in video activity reports. The second is to investigate the relationship between each component (Vigor, Dedication, and Absorption) and engagement. Based on this relationship, it could be possible to identify the factors that allow the estimation of engagement with high accuracy. The third is analysis of video information from the video activity reports. Sanghvi et al. and Rajavenkatanarayanan et al. claimed that body movements and facial expressions from videos can be used to measure engagement [Sanghvi et al., 2011, Rajavenkatanarayanan et al., 2018]. The unusual movement observed by He et al. [He et al., 2022] can also be considered for estimating engagement. By analyzing video information, also becomes possible to examine the differences in media used to measure engagement, such as audio and video.

6 Conclusion

This study is for supporting remote PBL with particular focus on communication between a teacher and students engaging in remote PBL. To create a mechanism for a teacher to understand students' engagement, we analyzed the relationship between the video activity reports from students and theier engagement, which has not been sufficiently addressed.

From video activity reports submitted once a day from eight graduate students majoring in informatics, we found that engagement decreased (1) when the reports contained negative words, (2) when filled pauses were frequent or long, and (3) when silent pauses were infrequent or short. In addition, we developed a prototype system and discussed the design guidelines for a system using AI technology to estimate engagement considering individual differences and to assist in understanding learner engagement.

References

[Adnan and Anwar, 2020] Adnan, M. and Anwar, K. (2020). Online Learning amid the COVID-19 Pandemic: Students' Perspectives. *Online Submission*, 2(1):45–51.

[Afflerbach and Harrison, 2017] Afflerbach, P. and Harrison, C. (2017). What Is Engagement, How Is It Different From Motivation, and How Can I Promote It? *Journal of Adolescent & Adult Literacy*, 61(2):217–220.

[Azevedo, 2015] Azevedo, R. (2015). Defining and Measuring Engagement and Learning in Science: Conceptual, Theoretical, Methodological, and Analytical Issues. *Educational Psychologist*, 50(1):84–94.

[Bailey and Kurland, 2002] Bailey, D. E. and Kurland, N. B. (2002). A Review of Telework Research: Findings, New Directions, and Lessons for the Study of Modern Work. *Journal of Organizational Behavior: The International Journal of Industrial, Occupational and Organizational Psychology and Behavior*, 23(4):383–400.

[Bakker and Demerouti, 2008] Bakker, A. B. and Demerouti, E. (2008). Towards a Model of Work Engagement. *Career development international*, 13(3):209–223.

[Baruch, 2000] Baruch, Y. (2000). Teleworking: Benefits and Pitfalls as Perceived by Professionals and Managers. *New Technology, Work and Employment*, 15(1):34–49.

[Beckel and Fisher, 2022] Beckel, J. L. and Fisher, G. G. (2022). Telework and Worker Health and Well-Being: A Review and Recommendations for Research and Practice. *International Journal of Environmental Research and Public Health*, 19(7):3879.

[Benke et al., 2020a] Benke, I., Knierim, M. T., and Maedche, A. (2020a). Chatbot-Based Emotion Management for Distributed Teams: A Participatory Design Study. *Proc. ACM Hum.-Comput. Interact.*, 4(CSCW2).

[Benke et al., 2020b] Benke, I., Knierim, M. T., and Maedche, A. (2020b). Chatbot-based Emotion Management for Distributed Teams: A Participatory Design Study. *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW2):1–30.

[Bezerra et al., 2020] Bezerra, C. I. M., de Souza Filho, J. C., Coutinho, E. F., Gama, A., Ferreira, A. L., de Andrade, G. L. a., and Feitosa, C. E. (2020). How Human and Organizational Factors Influence Software Teams Productivity in COVID-19 Pandemic: A Brazilian Survey. In *Proceedings of the 34th Brazilian Symposium on Software Engineering*, SBES '20, pages 606–615, New York, NY, USA. Association for Computing Machinery.

[Breevaart et al., 2012] Breevaart, K., Demerouti, E., and Hetland, J. (2012). The Measurement of State Work Engagement A Multilevel Factor Analytic Study. *European Journal of Psychological Assessment*, 28:305.

[Breideband et al., 2022] Breideband, T., Talkad Sukumar, P., Mark, G., Caruso, M., D'Mello, S., and Striegel, A. D. (2022). Home-Life and Work Rhythm Diversity in Distributed Teamwork: A Study with Information Workers during the COVID-19 Pandemic. *Proceedings of the ACM on Human-Computer Interaction*, 6(CSCW1):1–23.

[Brennan and Williams, 1995] Brennan, S. E. and Williams, M. (1995). The Feeling of Another's Knowing: Prosody and Filled Pauses as Cues to Listeners about the Metacognitive States of Speakers. *Journal of memory and language*, 34(3):383–398.

[Buchanan et al., 2014] Buchanan, T. W., Laures-Gore, J. S., and Duff, M. C. (2014). Acute Stress Reduces Speech Fluency. *Biological Psychology*, 97:60–66.

[Bänziger et al., 2009] Bänziger, T., Grandjean, D., and Scherer, K. (2009). Emotion Recognition From Expressions in Face, Voice, and Body: The Multimodal Emotion Recognition Test (MERT). *Emotion (Washington, D.C.)*, 9:691–704.

[Campione and Véronis, 2002] Campione, E. and Véronis, J. (2002). A Large-Scale Multilingual Study of Silent Pause Duration. In *Speech prosody 2002, international conference*.

[Cao et al., 2021] Cao, H., Yang, V., Chen, V., Lee, Y. J., Stone, L., Diarrassouba, N. J., Whiting, M. E., and Bernstein, M. S. (2021). My Team Will Go On: Differentiating High and Low Viability Teams through Team Interaction. *Proc. ACM Hum.-Comput. Interact.*, 4(CSCW3).

[Chen et al., 2021] Chen, Z., Cao, H., Deng, Y., Gao, X., Piao, J., Xu, F., Zhang, Y., and Li, Y. (2021). Learning from Home: A Mixed-Methods Analysis of Live Streaming Based Remote Education Experience in Chinese Colleges during the COVID-19 Pandemic. In *Proceedings of the 2021 CHI Conference on human factors in computing systems*, pages 1–16.

[Chickering and Gamson, 1987] Chickering, A. W. and Gamson, Z. F. (1987). Seven Principles For Good Practice in Undergraduate Education. *AAHE bulletin*, 3:7.

[Chounta et al., 2017] Chounta, I.-A., Manske, S., and Hoppe, H. U. (2017). "From Making to Learning": Introducing Dev Camps As an Educational Paradigm for Re-Inventing Problem-Based Learning. *International Journal of Educational Technology in Higher Education*, 14(1):1–15.

[Christenfeld and Creager, 1996] Christenfeld, N. and Creager, B. (1996). Anxiety, Alcohol, Aphasia, and Ums. *Journal of personality and social psychology*, 70(3):451.

[Cojocariu et al., 2014] Cojocariu, V.-M., Lazar, I., Nedeff, V., and Lazar, G. (2014). SWOT Anlysis of E-learning Educational Services from the Perspective of their Beneficiaries. *Procedia-Social and Behavioral Sciences*, 116:1999–2003.

[de Souza Santos and Ralph, 2022] de Souza Santos, R. E. and Ralph, P. (2022). A Grounded Theory of Coordination in Remote-First and Hybrid Software Teams. In *Proceedings of the 44th International Conference on Software Engineering*, ICSE '22, pages 25–35, New York, NY, USA. Association for Computing Machinery.

[De Vincenzi et al., 2022] De Vincenzi, C., Pansini, M., Ferrara, B., Buonomo, I., and Benevene, P. (2022). Consequences of COVID-19 on Employees in Remote Working: Challenges, Risks and Opportunities An Evidence-Based Literature Review. *International Journal of Environmental Research and Public Health*, 19(18):11672.

[Delaborde and Devillers, 2010] Delaborde, A. and Devillers, L. (2010). Use of Nonverbal Speech Cues in Social Interaction between Human and Robot: Emotional and Interactional Markers. In *Proceedings of the 3rd International Workshop on Affective Interaction in Natural Environments*, AFFINE '10, page 75 – 80, New York, NY, USA. Association for Computing Machinery.

[Demerouti et al., 2001] Demerouti, E., Nachreiner, F., and Schaufeli, W. (2001). The Job Demands – Resources Model of Burnout. *The Journal of applied psychology*, 86:499–512.

[Devlin et al., 2019] Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

[Dimitriadou et al., 2020] Dimitriadou, S., Lavidas, K., Karalis, T., and Ravanis, K. (2020). Study Engagement in University Students: a Confirmatory Factor Analysis of the Utrecht Work Engagement Scale with Greek Students. *Journal of Well-Being Assessment*, 4(3):291–307.

[Duez, 1982] Duez, D. (1982). Silent and Non-Silent Pauses in Three Speech Styles. *Language and speech*, 25(1):11–28.

[Duez, 1985] Duez, D. (1985). Perception of Silent Pauses in Continuous Speech. *Language and speech*, 28(4):377–389.

[El-Shinnawy and Markus, 1997] El-Shinnawy, M. and Markus, M. (1997). The Poverty of Media Richness Theory: Explaining People's Choice of Electronic Mail Vs. Voice Mail. *International Journal of Human-Computer Studies*, 46(4):443–467.

[Etkina and Harper, 2002] Etkina, E. and Harper, K. A. (2002). Weekly Reports: Student Reflections on Learning. *Journal of College Science Teaching*, 31(7):476–480.

[Finn and Zimmer, 2012] Finn, J. D. and Zimmer, K. S. (2012). Student Engagement: What Is It? Why Does It Matter? In *Handbook of research on student engagement*, pages 97–131. Springer.

[Fujisaki, 2004] Fujisaki, H. (2004). Information, Prosody, and Modeling – with Emphasis on Tonal Features of Speech. *Scientific Programming - SP*.

[Galanti et al., 2021] Galanti, T., Guidetti, G., Mazzei, E., Zappalà, S., and Toscano, F. (2021). Work from Home during the Covid-19 Outbreak: The Impact on Employees' Remote Work Productivity, Engagement, and Stress. *Journal of occupational and environmental medicine*, 63(7):e426.

[Gilson et al., 2015] Gilson, L. L., Maynard, M. T., Jones Young, N. C., Vartiainen, M., and Hakonen, M. (2015). Virtual Teams Research: 10 Years, 10 Themes, and 10 Opportunities. *Journal of management*, 41(5):1313–1337.

[Golden, 2009] Golden, T. D. (2009). Applying Technology to Work: Toward a Better Understanding of Telework. *Organization Management Journal*, 6(4):241–250.

[Golestani et al., 2018] Golestani, A., Masli, M., Shami, N. S., Jones, J., Menon, A., and Mondal, J. (2018). Real-Time Prediction of Employee Engagement Using Social Media and Text Mining. In *2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pages 1383–1387.

[Goto et al., 1999] Goto, M., Itou, K., and Hayamizu, S. (1999). A Real-Time Filled Pause Detection System For Spontaneous Speech Recognition. In *Sixth European Conference on Speech Communication and Technology*.

[Harter et al., 2013] Harter, J. K., Schmidt, F. L., Agrawal, S., Plowman, S. K., and Blue, A. (2013). The Relationship Between Engagement at Work and Organizational Outcomes. *Gallup Poll Consulting University Press, Washington*.

[Harter et al., 2002] Harter, J. K., Schmidt, F. L., and Hayes, T. L. (2002). Business-Unit-Level Relationship Between Employee Satisfaction, Employee Engagement, and Business Outcomes: A Meta-Analysis. *Journal of applied psychology*, 87(2):268.

[Hassib et al., 2017] Hassib, M., Buschek, D., Wozniak, P. W., and Alt, F. (2017). HeartChat: Heart Rate Augmented Mobile Chat to Support Empathy and Awareness. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, pages 2239–2251.

[He, 2022] He, Z. (2022). テレワーク環境におけるワークエンゲージメント向上のための活動報告の研究. Master's thesis, University of Tsukuba. (in Japanese).

[He et al., 2020a] He, Z., Dai, X., Yamakami, T., and Inoue, T. (2020a). Preliminary Utility Study of a Short Video as a Daily Report in Teleworking. In *International Conference on Collaboration Technologies and Social Computing*, pages 35–49. Springer.

[He et al., 2020b] He, Z., Dai, X., Yamakami, T., Kubota, S., and Inoue, T. (2020b). テレワーク環境における従業員エンゲージメントのための動画日報システムの提案. In *IEICE Tech. Rep.*, volume 120, pages 40–45. (in Japanese).

[He et al., 2022] He, Z., Sarcar, S., and Tomoo, I. (2022). Exploring the Feasibility of Video Activity Reporting for Students in Distance Learning. In *Data Science, Human-Centered Computing, and Intelligent Technologies*, pages 44–55.

[Hu et al., 2022] Hu, X. E., Hinds, R., Valentine, M., and Bernstein, M. S. (2022). A "Distance Matters" Paradox: Facilitating Intra-Team Collaboration Can Harm Inter-Team Collaboration. *Proceedings of the ACM on Human-Computer Interaction*, 6(CSCW1):1–36.

[Ito et al., 2016] Ito, K., Kizuka, A., and Oba, M. (2016). A Trial Utilization of Weekly Reports to Evaluate Learning for System Development PBLs. In *2016 5th IIAI International Congress on Advanced Applied Informatics (IIAI-AAI)*, pages 1064–1067. IEEE.

[Johar, 2014] Johar, S. (2014). Paralinguistic Profiling Using Speech Recognition. *International Journal of Speech Technology*, 17:205–209.

[Johnson et al., 2021] Johnson, B., Zimmermann, T., and Bird, C. (2021). The Effect of Work Environments on Productivity and Satisfaction of Software Engineers. *IEEE Transactions on Software Engineering*, 47:736–757.

[Kahn et al., 2007] Kahn, J. H., Tobin, R. M., Massey, A. E., and Anderson, J. A. (2007). Measuring Emotional Expression with the Linguistic Inquiry and Word Count. *The American journal of psychology*, 120(2):263–286.

[Kahn, 1990] Kahn, W. A. (1990). Psychological Conditions of Personal Engagement and Disengagement at Work. *Academy of management journal*, 33(4):692–724.

[Kajiwara et al., 2019] Kajiwara, Y., Shimauchi, T., and Kimura, H. (2019). Predicting Emotion and Engagement of Workers in Order Picking Based on Behavior and Pulse Waves Acquired by Wearable Devices. *Sensors*, 19(1):165.

[Lee et al., 2020] Lee, I. C. J., Koh, H., Lai, S. H., and Hwang, N. C. (2020). Academic coaching of medical students during the COVID-19 pandemic. *Medical Education*, 54(12):1184–1185.

[Lee et al., 2017] Lee, M., Kim, J., Truong, K., de Kort, Y., Beute, F., and IJsselsteijn, W. (2017). Exploring Moral Conflicts in Speech: Multidisciplinary Analysis of Affect and Stress. In *2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 407–414.

[Longhurst et al., 2020] Longhurst, G. J., Stone, D. M., Dulohery, K., Scully, D., Campbell, T., and Smith, C. F. (2020). Strength, Weakness, Opportunity, Threat (SWOT) Analysis of the Adaptations to Anatomical Education in the United Kingdom and Republic of Ireland in Response to the Covid-19 Pandemic. *Anatomical sciences education*, 13(3):301–311.

[Loscalzo and Giannini, 2019] Loscalzo, Y. and Giannini, M. (2019). Study Engagement in Italian University Students: A Confirmatory Factor Analysis of the Utrecht Work Engagement Scale-Student Version. *Social Indicators Research*, 142(2):845–854.

[Lu et al., 2018] Lu, D., Marlow, J., Kocielnik, R., and Avrahami, D. (2018). Challenges and Opportunities for Technology-Supported Activity Reporting in the Workplace. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, pages 1–12.

[Markham, 2011] Markham, T. (2011). Project Based Learning A Bridge Just Far Enough. *Teacher Librarian*, 39(2):38–42.

[Marks, 2000] Marks, H. M. (2000). Student Engagement in Instructional Activity: Patterns in the Elementary, Middle, and High School Years. *American educational research journal*, 37(1):153–184.

[McGill et al., 2019] McGill, M. M., Decker, A., McKlin, T., and Haynie, K. (2019). A Gap Analysis of Noncognitive Constructs in Evaluation Instruments Designed for Computing Education. In *Proceedings of the 50th ACM Technical Symposium on Computer Science Education*, SIGCSE '19, pages 706–712, New York, NY, USA. Association for Computing Machinery.

[McManus and Costello, 2019] McManus, J. W. and Costello, P. J. (2019). Project Based Learning in Computer Science: A Student and Research Advisor's Perspective. *J. Comput. Sci. Coll.*, 34(3):38 – 46.

[Mitra et al., 2017] Mitra, T., Muller, M., Shami, N. S., Golestani, A., and Masli, M. (2017). Spread of Employee Engagement in a Large Organizational Network: A Longitudinal Analysis. *Proc. ACM Hum.-Comput. Interact.*, 1(CSCW).

[Mulki et al., 2009] Mulki, J. P., Bardhi, F., Lassk, F. G., and Nanavaty-Dahl, J. (2009). Set Up Remote Workers to Thrive. *MIT Sloan Management Review*, 51(1):63.

[Nguyen et al., 2022] Nguyen, M., Laly, M., Kwon, B. C., Mougenot, C., and McNamara, J. (2022). Moody Man: Improving Creative Teamwork through Dynamic Affective Recognition. In *Extended Abstracts of the 2022 CHI Conference on Human Factors in Computing Systems*, CHI EA '22, New York, NY, USA. Association for Computing Machinery.

[Olson, 1983] Olson, M. H. (1983). Remote Office Work: Changing Work Patterns in Space and Time. *Communications of the ACM*, 26(3):182–187.

[Overmyer, 2011] Overmyer, S. P. (2011). *Implementing Telework: Lessons Learned from Four Federal Agencies*. IBM Center for the Business of Government Arlington, VA.

[Pearlson and Saunders, 2001] Pearlson, K. and Saunders, C. (2001). There's No Place Like Home: Managing Telecommuting Paradoxes. *Academy of Management Executive*, 15:117–128.

[Pogorilich, 1992] Pogorilich, D. A. (1992). The Daily Report as a Job Management Tool: a Publication of the American Association of Cost Engineers. *Cost Engineering*, 34(2):23.

[Pope et al., 1970] Pope, B., Blass, T., Siegman, A. W., and Raher, J. (1970). Anxiety and Depression in Speech. *Journal of Consulting and Clinical Psychology*, 35(1p1):128.

[Pyöriä, 2011] Pyöriä, P. (2011). Managing Telework: Risks, Fears and Rules. *Management Research Review*, 34:386–399.

[Rajavenkatanarayanan et al., 2018] Rajavenkatanarayanan, A., Babu, A. R., Tsiakas, K., and Makedon, F. (2018). Monitoring Task Engagement Using Facial Expressions and Body Postures. In *Proceedings of the 3rd International Workshop on Interactive and Spatial Computing*, IWISC '18, page 103 – 108, New York, NY, USA. Association for Computing Machinery.

[Rochester, 1973] Rochester, S. R. (1973). The Significance of Pauses in Spontaneous Speech. *Journal of Psycholinguistic Research*, 2:51–81.

[Rumberger, 1987] Rumberger, R. W. (1987). High School Dropouts: A Review of Issues and Evidence. *Review of educational research*, 57(2):101–121.

[Sanghvi et al., 2011] Sanghvi, J., Castellano, G., Leite, I., Pereira, A., McOwan, P. W., and Paiva, A. (2011). Automatic Analysis of Affective Postures and Body Motion to Detect Engagement with a Game Companion. In *Proceedings of the 6th International Conference on Human-Robot Interaction*, HRI '11, page 305 – 312, New York, NY, USA. Association for Computing Machinery.

[Sarker et al., 2012] Sarker, S., Sarker, S., Xiao, X., and Ahuja, M. (2012). Managing Employees' Use of Mobile Technologies to Minimize Work-Life Balance Impacts. *MIS Quarterly Executive*, 11:143–157.

[Schaufeli and Bakker, 2004] Schaufeli, W. B. and Bakker, A. B. (2004). Job Demands, Job Resources, and Their Relationship with Burnout and Engagement: A Multi-Sample Study. *Journal of Organizational Behavior: The International Journal of Industrial, Occupational and Organizational Psychology and Behavior*, 25(3):293–315.

[Schaufeli et al., 2006] Schaufeli, W. B., Bakker, A. B., and Salanova, M. (2006). The Measurement of Work Engagement With a Short Questionnaire: A Cross-National Study. *Educational and psychological measurement*, 66(4):701–716.

[Schaufeli et al., 2002a] Schaufeli, W. B., Martinez, I. M., Pinto, A. M., Salanova, M., and Bakker, A. B. (2002a). Burnout and Engagement in University Students: A Cross-National Study. *Journal of cross-cultural psychology*, 33(5):464–481.

[Schaufeli et al., 2002b] Schaufeli, W. B., Salanova, M., González-Romá, V., and Bakker, A. B. (2002b). The Measurement of Engagement and Burnout: A Two Sample Confirmatory Factor Analytic Approach. *Journal of Happiness studies*, 3(1):71–92.

[Schaufeli et al., 2017] Schaufeli, W. B., Shimazu, A., Hakanen, J., Salanova, M., and De Witte, H. (2017). An Ultra-Short Measure for Work Engagement. *European Journal of Psychological Assessment*.

[Shami et al., 2015] Shami, N. S., Muller, M., Pal, A., Masli, M., and Geyer, W. (2015). Inferring Employee Engagement from Social Media. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, CHI '15, pages 3999–4008, New York, NY, USA. Association for Computing Machinery.

[Sonnentag et al., 2010] Sonnentag, S., Dormann, C., Demerouti, E., et al. (2010). Not All Days Are Created Equal: The Concept of State Work Engagement. *Work engagement: A handbook of essential theory and research*, pages 25–38.

[Suh, 1999] Suh, K. S. (1999). Impact of communication medium on task performance and satisfaction: an examination of media-richness theory. *Information & Management*, 35(5):295–312.

[Tanaka et al., 2021] Tanaka, H., Yamada, W., and Ochiai, K. (2021). Estimating Work Engagement from Online Chat Logs. In *Asian CHI Symposium 2021*, Asian CHI Symposium 2021, pages 70–73, New York, NY, USA. Association for Computing Machinery.

[Tang et al., 2022] Tang, Q., Hu, X., Zeng, Z., and Zhao, Y. (2022). Co-Orb: Fostering Remote Workplace Gratitude with IoT Technology. In *Extended Abstracts of the 2022 CHI Conference on Human Factors in Computing Systems*, CHI EA '22, New York, NY, USA. Association for Computing Machinery.

[Tremblay and Thomsin, 2012] Tremblay, D.-G. and Thomsin, L. (2012). Telework and Mobile Working: Analysis of Its Benefits and Drawbacks. *Int. J. of Work Innovation*, 1:100 – 113.

[Trowler, 2010] Trowler, V. (2010). Student Engagement Literature Review. *The higher education academy*, 11(1):1–15.

[Veinott et al., 1999] Veinott, E. S., Olson, J., Olson, G. M., and Fu, X. (1999). Video Helps Remote Work: Speakers Who Need to Negotiate Common Ground Benefit from Seeing Each Other. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '99, page 302 – 309, New York, NY, USA. Association for Computing Machinery.

[Wang et al., 2022] Wang, D., Wang, H., Yu, M., Ashktorab, Z., and Tan, M. (2022). Group Chat Ecology in Enterprise Instant Messaging: How Employees Collaborate Through Multi-User Chat Channels on Slack. *Proceedings of the ACM on Human-Computer Interaction*, 6(CSCW1):1–14.

[Wang and Hu, 2018] Wang, Y. and Hu, W. (2018). Speech Emotion Recognition Based on Improved MFCC. In *Proceedings of the 2nd International Conference on Computer Science and Application Engineering*, CSAE '18, New York, NY, USA. Association for Computing Machinery.

[Wilcha et al., 2020] Wilcha, R.-J. et al. (2020). Effectiveness of Virtual Medical Teaching During the COVID-19 Crisis: Systematic Review. *JMIR medical education*, 6(2):e20963.

[Zhuang and Chen, 2020] Zhuang, J. and Chen, Y.-W. (2020). Research on the Impact of Employee Growth on Work Performance from the Perspective of Positive Psychology. In *2020 The 11th International Conference on E-Business, Management and Economics*, ICEME 2020, pages 201–205, New York, NY, USA. Association for Computing Machinery.

Journal of Universal Computer Science, vol. 29, no. 11 (2023), 1361-1384
submitted: 19/5/2023, accepted: 30/9/2023, appeared: 28/11/2023 CC BY-ND 4.0

Towards a Traceable Data Model Accommodating Bounded Uncertainty for DST Based Computation of *BRCA1/2* Mutation Probability With Age

Lorenz Gillner
(University of Applied Sciences Wismar, Germany
https://orcid.org/0009-0007-8244-5810, lorenz.gillner@hs-wismar.de)

Ekaterina Auer
(University of Applied Sciences Wismar, Germany
https://orcid.org/0000-0003-4059-3982, ekaterina.auer@hs-wismar.de)

Abstract: In this paper, we describe the requirements for traceable open-source data retrieval in the context of computation of *BRCA1/2* mutation probabilities (mutations in two tumor-suppressor genes responsible for hereditary BReast or/and ovarian CAncer). We show how such data can be used to develop a Dempster-Shafer model for computing the probability of *BRCA1/2* mutations enhanced by taking into account the actual age of a patient or a family member in an appropriate way even if it is not known exactly. The model is compared with PENN II and BOADICEA (based on undisclosed data), two established platforms for this purpose accessible online, as well as with our own previous models. A proof-of-concept implementation shows that set-based techniques are able to provide better information about mutation probabilities, simultaneously highlighting the necessity for ground truth data of high quality.

Keywords: HBOC, *BRCA1/2* , mutation probability, data integration, data fusion, interval analysis, Dempster-Shafer evidence theory, provenance, AI based data mining
Categories: B.8, E.2, G.1.0, G.1.10, G.3, H.3, H.4.2, I.6.4, J.3
DOI: 10.3897/jucs.112797

1 Introduction

Pathogenic variants (also known as mutations) in cells are responsible for human diseases, most notably cancer. The majority of cancer cases are caused by the so-called somatic mutations, that is, changes in the DNA sequence of non-reproductive cells. However, if such changes occur in the reproductive cells (germline mutations), they can be inherited/passed on to the next generation, which accounts for a high cancer risk not only for specific persons but for their entire families. In particular, variants occurring in *BRCA1* and *BRCA2* genes are the most well-known mutations leading to hereditary breast cancer (BC). In comparison to approximately 13% risk of developing BC during a lifetime in the general female population, the risk of developing BC by 70 – 80 years of age is increased to 55% – 72% if there are hereditary mutations in *BRCA1* and to 45% – 69% if a hereditary *BRCA2* mutation is detected[1]. Additionally, the likelihood of contracting ovarian cancer (OC) becomes higher. The corresponding phenomenon is called the hereditary BC and OC (HBOC) syndrome[2].

[1] https://www.cancer.gov/about-cancer/causes-prevention/genetics/brca-fact-sheet

[2] To avoid the implication that this syndrome affects only women, scientists also start to name it King syndrome.

Since there is a strong correlation between HBOC and *BRCA1/2* mutations, a possible strategy for identifying persons or families at high risk is to take a look at whether they carry a corresponding pathogenic variant. That is the reason why genetic testing and counseling for such kinds of mutations gain more and more importance nowadays. Although actual diagnostic genetic testing for DNA changes is getting cheaper, it still cannot be recommended for everyone. Therefore, mutation probability prediction software is used in genetic counseling (investigating if patients have a hereditary risk of a disease). Such software attempts to compute mutation probabilities for patients based on specific indicators (e.g., their family history, ethnicity or origin) without the actual process of testing. At the moment, there exist several established risk assessment (RA) tools and questionnaires for the purpose of predicting mutation or cancer risk, some of them accessible online (e.g., PENN II[3] or BOADICEA[4]). However, such tools are often based on undisclosed or untraceable data leading to questions about validity areas and credibility of results. For example, risks computed by PENN II for a *BRCA1/2* mutation are occasionally quite different from BOADICEA for the same person, without the possibility for non-expert patients to check or understand why or even to decide what prediction to trust (cf. Section 2.3). In [Auer and Luther 2021], we provided an overview of the state of the art, HBOC-related (meta)studies, tools and questionnaires and pointed out further possible problems with their use.

One of the first challenges for a research team trying to develop a new model for *BRCA1/2* based on traceable data is to obtain ground truth information. Medical professionals with access to large amounts of data from clinical trials publish statistical results, but the exact composition of the study population or models used for estimates often remain unclear. Genetic research organizations might disclose the algorithms used to generate a model, but its factual basis remains inaccessible to outsiders [Guerrini et al. 2017]. Although many publications in the area of medical science are open access (OA), the relevant source material is not. All this is justified by data protection legislation: A person's DNA constitutes sensitive medical information since it is a biometric trait. Therefore, it must be handled with utmost care to prevent invasion of individuals' privacy or identity theft.

This leads to a dilemma within genetic research since the correlation between certain types of cancer and the genetic profiles of patients is often exactly the subject of interest. Studying it requires access to confidential records, often obtained in collaboration with specialized institutions. A few projects dedicated to the collection of anonymized clinical data are publicly available over the internet. Two such repositories are the GDC[5] and ICGC[6] data portals. They allow users to explore data sets on various cancer cases from all over the world, with supplemental molecular samples including mutations. However, those projects severely restrict the access to data on germline variants by the requirements of a rigorous registration process. Simple germline variants (SGVs) differ from simple somatic mutations (SSMs) in that they appear in every DNA sample of a donor (and possibly their relatives), which can lead to the identification of the individual in question. However, for modeling the risk of carrying an inherited pathogenic *BRCA1/2* variant, an opportunity to use SSMs does not suffice.

Aside from the data access hurdle, a big difficulty in devising appropriate risk assessment tools in the context of HBOC is the uncertainty in the underlying data, its

[3] https://pennmodel2.pmacs.upenn.edu/penn2/

[4] https://ccge.medschl.cam.ac.uk/boadicea/

[5] https://gdc.cancer.gov/

[6] https://dcc.icgc.org/

major (interconnected) sources being data provenance and collection. A good record of data provenance facilitates reproducibility, validation and belief in the reliability of scientific results [Pasquier et al. 2017]. This point, especially important in the area of medical science, is still not implemented to a sufficient degree there. For example, although GDC contains data base items about the specific history of a family or patient connected to a certain variant (e.g., ethnicity, origin, age), they are often left empty (at least, in the publicly accessible somatic case). Frequently, the way of collecting the data is not documented sufficiently. For example, cohort composition or exact criteria for the choice of test persons as often as not remain unclear within a survey. Besides, there is inherent uncertainty present in the data since patients are often unsure about the specifics of their family history. Although it is difficult to remove this latter cause of uncertainty, the reduction in the former two can and should be addressed by astute researchers.

Uncertainty is present in practically every kind of real-life application, but it is especially high in the case of medical studies. In the field of uncertainty quantification, two main sources are discerned: aleatory (due to randomness) and epistemic (due to the lack of knowledge). State of the art tools such as PENN II or BOADICEA rely on crisp data in combination with the classical probability theory to take into account aleatory uncertainty, working with arithmetic means or medians in the presence of the epistemic one. In many cases, intervals represent the available, uncertain but bounded data better than crisp numbers. They can be propagated from inputs to outputs of a (static or dynamic) model using interval analysis (IA) [Moore et al. 2009]. Moreover, a way of combining the probabilistic and set-based reasoning is offered by the co-called imprecise probability [Bradley 2019], in particular, the Dempster-Shafer evidence theory (DST) [Shafer 1976].

The goals of this paper are, first, to formulate the requirements on the optimal HBOC database from the viewpoint of risk assessment with a focus on reliability including traceability and provenance. Second, we propose an interval DST model for computing *BRCA1/2* mutation probabilities based on data from OA publications as a proof of concept that taking into account epistemic uncertainty explicitly provides improved quality information for a patient. This model helps to incorporate uncertainty about the ages of the involved persons better. We compare the results with those from PENN II and BOADICEA as well as with those from our earlier model that did not differentiate to the same degree wrt. age. Here, the importance of careful extraction of ground truth data can be seen: we provide results of the same (old) model from [Auer and Luther 2021] based solely on data from [Frank et al. 2002] versus aggregated data extracted from several OA publications. The data in the developed database is traceable to the publication where they were provided.

The paper is structured as follows. First, we provide a short overview of the background methods and tools we rely on in Section 2. After that, we formulate the requirements on the (ideal) HBOC data base and provide a proof-of concept implementation filled by data from OA publications in Section 3. Next, we present our two-phase DST model improving that from [Auer and Luther 2022] by taking into account the uncertainty about age using intervals in Section 4. Conclusions and an outlook on our future work are in the last section.

2 Background

In this section, interval analysis and the Dempster-Shafer theory are described briefly. Additionally, we overview the main features of PENN II and BOADICEA, two models based on conventional probability theory and available online for computing the *BRCA1/2* mutation probability and the risk of breast cancer, respectively.

2.1 Basic concepts of interval analysis

Interval analysis [Moore et al. 2009] is a well-known tool for result verification with applications in computer-assisted proofs, engineering, computer graphics, medical science and many others. With the help of IA, it is possible to prove formally, using a suitable fixed point theorem the assumptions of which can be checked on a computer reliably, that the result of a computer simulation is correct (given that the underlying code is correct). This takes into account such factors as rounding, conversion, discretization or truncation errors. The results are intervals with bounds expressed by floating point numbers which with certainty contain the exact solution to the formal model. Since the methods work with sets, they can be used for propagating bounded uncertainty, usually from the inputs to the outputs, in a deterministic way although there are also approaches to solve inverse propagation problems [Merheb et al. 2013, Desrochers and Jaulin 2017]. A common drawback of such rigor-preserving methods, caused by the dependency problem or the wrapping effect [Lohner 2001], is the possibility of too wide solution sets – an inherent problem of naive IA that more sophisticated techniques with result verification (e.g., affine or Taylor model based approaches) address [de Figueiredo and Stolfi 2004, Neumaier 2003, Makino and Berz 2004].

A real interval $[\underline{x}, \overline{x}]$, where $\underline{x} \in \mathbb{R}$ is the lower, $\overline{x} \in \mathbb{R}$ the upper bound, is defined as

$$[\underline{x}, \overline{x}] = \{x \in \mathbb{R} | \underline{x} \leq x \leq \overline{x}\} \ ,$$

usually for $\underline{x} \leq \overline{x}$. Crisp numbers $x \in \mathbb{R}$ can be represented by point intervals with $\underline{x} = \overline{x} = x$. For an operation $\circ = \{+, -, \cdot, /\}$ and two intervals $[\underline{x}, \overline{x}]$, $[\underline{y}, \overline{y}]$, the corresponding interval operation can be defined as

$$\begin{aligned}[\underline{x}, \overline{x}] \circ [\underline{y}, \overline{y}] :=& \{x \circ y \mid \forall x \in [\underline{x}, \overline{x}], y \in [\underline{y}, \overline{y}]\} \\ =& [\min(\underline{x} \circ \underline{y}, \underline{x} \circ \overline{y}, \overline{x} \circ \underline{y}, \overline{x} \circ \overline{y}), \max(\underline{x} \circ \underline{y}, \underline{x} \circ \overline{y}, \overline{x} \circ \underline{y}, \overline{x} \circ \overline{y})] \ ,\end{aligned}$$

that is, the result of an interval operation is also an interval. For normal interval division, it is assumed that $0 \notin [\underline{y}, \overline{y}]$ although it is possible to allow divisor intervals to contain zero in extended interval arithmetics (see, e.g., [Kahan 1968]). The general formula can be simplified for a given operation $\circ$ (e.g., $[\underline{x}, \overline{x}] - [\underline{y}, \overline{y}] = [\underline{x} - \overline{y}, \overline{x} - \underline{y}]$). An interval with floating point numbers as bounds can be obtained for any real interval in a verified way by using the concept of outward rounding. Based on interval arithmetic described above, which includes the possibility to evaluate functions over intervals, higher-level methods, for example, for solving systems of algebraic or differential equations, can be formulated to provide their error bounds (i.e., result verification) automatically.

2.2 Basic concepts of the Dempster-Shafer theory

The Dempster-Shafer theory [Ayyub and Klir 2006] facilitates synthesis between data and information and is increasingly used for uncertain data fusion, especially in the context of AI systems [Tang et al. 2023]. It combines evidence from different sources and provides a measure of confidence that a certain event occurs. A classical, additive, discrete probability density function defines the probability (given by a crisp number) that a random variable X is equal to its certain realization x_i (again, a crisp number, i.e., real or point value). The finite DST allows us to assign a (crisp) probability to the event that a realization of X belongs to a given set (e.g., an interval $[\underline{x}_i, \overline{x}_i]$)[7]. The result is given in terms of the lower and upper limits (belief and plausibility) on the probability of a subset of the frame of discernment Ω. A random DST variable can be characterized by its basic probability assignment (BPA) m. If $A_1, \ldots, A_n$ are the sets of interest where each $A_i \in 2^\Omega$, then m is defined by

$$\begin{aligned} m : 2^\Omega \to [0,1], \quad & m(A_i) = p_i,\ i = 1 \ldots n, \\ m(\emptyset) = 0, \quad & \sum_{i=1}^{n} m(A_i) = 1 \ . \end{aligned} \tag{1}$$

The mass of the impossible event $\emptyset$ is equal to zero. Every A_i with $m(A_i) \neq 0$ is called a focal element. The sum of masses of focal elements should be equal to one. If the sum is greater than one in BPAs provided by the experts, then a normalization can be carried out as $\tilde{m}(A_i) := m(A_i)/\sum\limits_{i=1}^{n} m(A_i)$. If the sum is less than one, then the same normalization can be used or a new focal element $A_{n+1} = \Omega$ can be introduced to accommodate the missing probability. The latter variant only makes sense for computing the lower limit $Bel(Y)$ whereas the former variant could inflate the belief function too much.

The plausibility ('worst case') and belief ('best case') functions can be defined with the help of the BPAs for all $i = 1 \ldots n$ and any $Y \subseteq \Omega$ as

$$Pl(Y) := \sum_{A_i \cap Y \neq \emptyset} m(A_i), \quad Bel(Y) := \sum_{A_i \subseteq Y} m(A_i). \tag{2}$$

These two functions represent a possibility to define an upper and a lower non-additive monotone measure [Ayyub and Klir 2006], respectively, on the true probability.

If there is evidence for the same issue from two or more sources, the BPAs have to be aggregated. In [Ferson et al. 2003], there is a good overview of the available aggregation methods, for example, Dempster's rule

$$m_{12}(A_i) = \frac{\sum\limits_{\forall A_j \cap A_k = A_i} m_1(A_j) m_2(A_k)}{1 - \sum\limits_{\forall A_j \cap A_k = \emptyset} m_1(A_j) m_2(A_k)} \tag{3}$$

with $A_i \neq \emptyset$, $m_{12}(\emptyset) = 0$ or mixing and averaging:

$$m_{1 \ldots n}(A_i) = \sum_{k=1}^{n} w_k \cdot m_k(A_i), \quad \sum_{k=1}^{n} w_k = 1 \ . \tag{4}$$

Although Dempster's rule in (3) is a fair way to combine conflict-free evidence, it cannot always be applied in the context of automatic data extraction since conflicts cannot be

[7] Analogous considerations can be made for continuous random variables

dismissed a priori there. Recently, a possibility to circumvent this has been proposed in [Tang et al. 2023].

As described, for example, in [Auer et al. 2010], it is possible to work with interval BPAs instead of crisp ones. The meaning of such an interval BPA (IBPA) is then as follows: The probability that a realization of a random variable X belongs to a certain set is itself uncertain (but bounded). The computation of $Pl(Y)$, $Bel(Y)$ and any kind of aggregation can work in the same way as for crisp BPAs if interval arithmetic is used instead of floating point arithmetic. Since we cannot define the inverse wrt. addition in interval arithmetic [Moore et al. 2009], the condition $\sum_{i=1}^{n} m(A_i) = 1$ for interval m cannot be fulfilled. It turns into the relaxed property $1 \in \sum_{i=1}^{n} m(A_i)$. The respective belief function signifying a lower limit on the strength of evidence is then itself an interval function having a lower and upper bound. Note that we do not try to countermeasure the relaxation of the summation property in this paper. It can be done in principle as shown, for example, in [Piegat and Dobryakova 2020] for linguistic probabilities (not for the DST) using interval arithmetic type 2. To use such kinds of IBPAs and compare the results to those presented in this paper is a topic of our future work.

2.3 Two existing web platforms for predicting *BRCA1/2* mutation probabilities

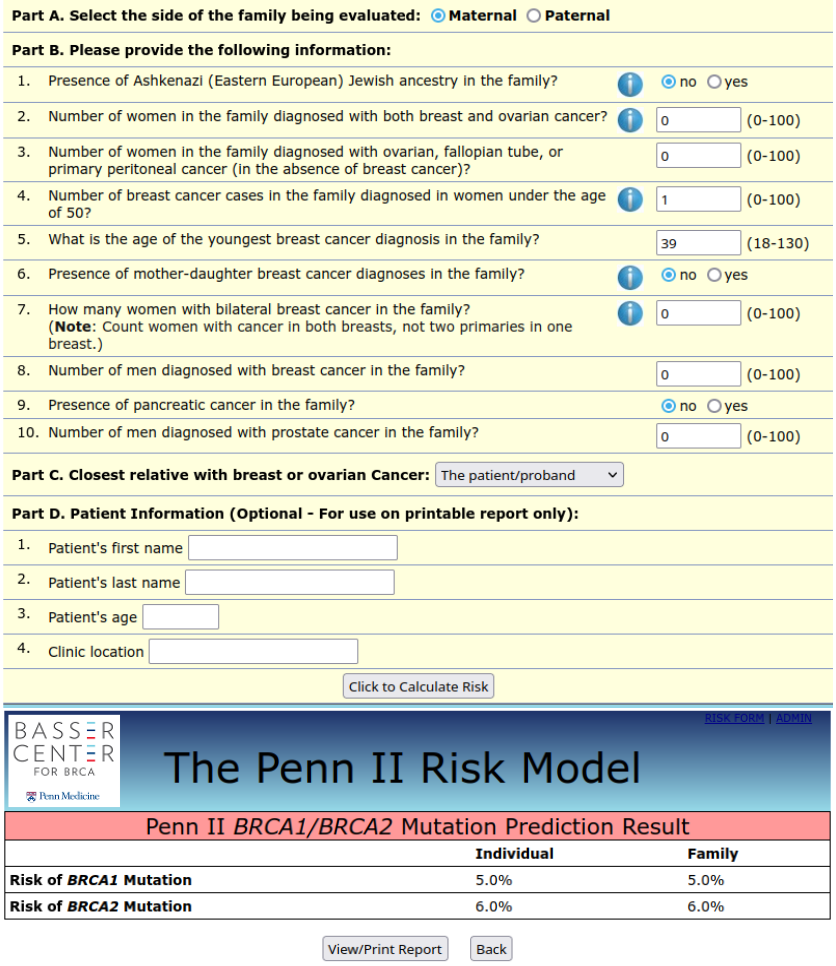
With over 8 million articles, PubMed® Central[8] is by far the most comprehensive public, English-language source for life science publications. More than half of the articles are available as OA. Nonetheless, or precisely because of this information flood, it is quite difficult for non-experts to find ground truth about the hereditary breast and ovarian cancer risk in their family.

In an early standard publication that had been used for prediction for many years, Tables from [Claus et al. 1994] estimate cumulative BC probability based on a survey considering mainly age-specific risk factors in combination with the family history and using a Bayesian model (with data on 4730 patients with confirmed BC matched against 4688 control subjects). Another relatively early paper [Frank et al. 2002] provides predictions for mutations in *BRCA1/2* correlated with such risk indicators as age of diagnosis, personal and family history, and ethnicity (also compiled in tables), for the cohort of overall 10000 participants (of Ashkenazi-Jewish and non-Ashkenazi-Jewish ethnicity). Frank tables could be seen as corresponding to ground truth since they contain observed frequencies. However, they are old, occasionally contradictory and consider only relatively small cohorts. The models PENN II and BOADICEA offer easy, questionnaire-type online interfaces to their respective models computing (among others) *BRCA1/2* mutation probabilities according to the risk indicators they consider important in a more detailed way. However, the actual data their models are based on are undisclosed.

PENN II [Lindor et al. 2010] is a mathematical model giving predictions about probabilities of *BRCA1* and *BRCA2* mutations based on logistic regression derived from 861 family histories of European and North American origin and taking into account Mendelian logic. The risk of a genetic defect is the same for the proband and the family if the proband is diagnosed with cancer. If not, the risk of the proband is reduced depending on the degree of relationship to the family member with a cancer diagnosis according to the principles of Mendelian genetics. This free tool offered by the University of Pennsylvania considers the following risk indicators: presence and ages of BC, presence of OC alone or with BC, bilaterality, diagnosis in both mother and daughter, male BC, presence of pancreatic and prostate cancers, Ashkenazi-Jewish or non-Ashkenazi-Jewish

[8] https://www.ncbi.nlm.nih.gov/pmc/

ethnicity as well as the degree of relation of the patient to the BC/OC case in the family. In Figure 1, screenshots of the PENN II questionnaire and result pages are shown for the web interface of the model.



Part A. Select the side of the family being evaluated: ● **Maternal** ○ **Paternal**

Part B. Please provide the following information:

1.	Presence of Ashkenazi (Eastern European) Jewish ancestry in the family?	● no ○ yes	
2.	Number of women in the family diagnosed with both breast and ovarian cancer?	0	(0-100)
3.	Number of women in the family diagnosed with ovarian, fallopian tube, or primary peritoneal cancer (in the absence of breast cancer)?	0	(0-100)
4.	Number of breast cancer cases in the family diagnosed in women under the age of 50?	1	(0-100)
5.	What is the age of the youngest breast cancer diagnosis in the family?	39	(18-130)
6.	Presence of mother-daughter breast cancer diagnoses in the family?	● no ○ yes	
7.	How many women with bilateral breast cancer in the family? (**Note**: Count women with cancer in both breasts, not two primaries in one breast.)	0	(0-100)
8.	Number of men diagnosed with breast cancer in the family?	0	(0-100)
9.	Presence of pancreatic cancer in the family?	● no ○ yes	
10.	Number of men diagnosed with prostate cancer in the family?	0	(0-100)

Part C. Closest relative with breast or ovarian Cancer: The patient/proband

Part D. Patient Information (Optional - For use on printable report only):

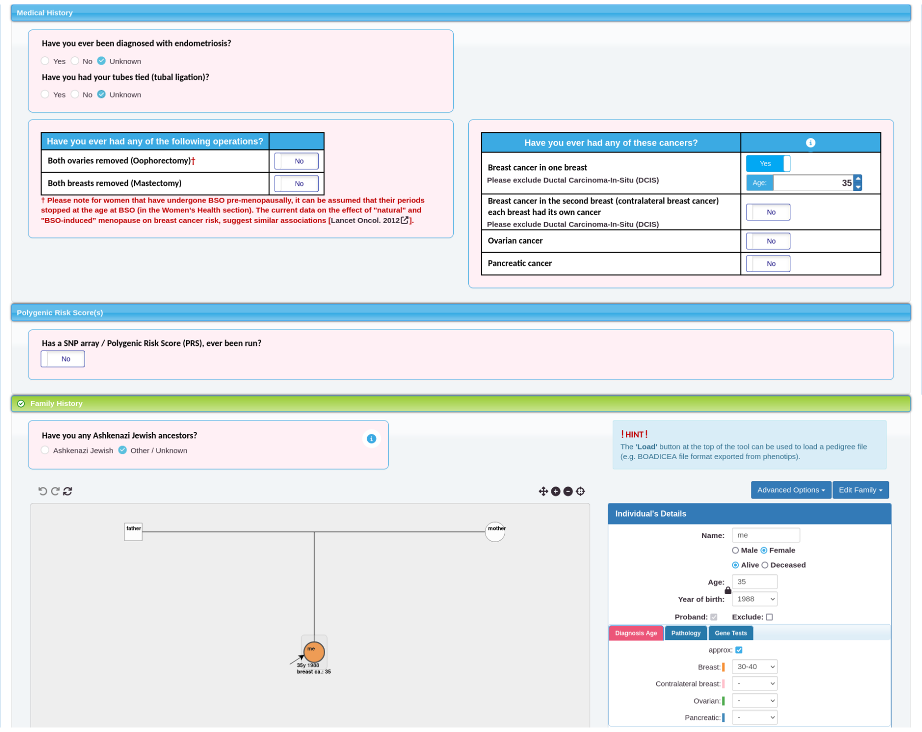
1. Patient's first name
2. Patient's last name
3. Patient's age
4. Clinic location

Penn II *BRCA1/BRCA2* Mutation Prediction Result

	Individual	**Family**
Risk of *BRCA1* Mutation	5.0%	5.0%
Risk of *BRCA2* Mutation	6.0%	6.0%

Figure 1: Screenshot of PENN II web interface from https://pennmodel2.pmacs.upenn.edu/penn2/: questions (above) and computed probabilities (below) for the disease pattern from this subsection

BOADICEA (Breast and Ovarian Analysis of Disease Incidence and Carrier Estimation Algorithm) [Lee et al. 2019] is a part of the originally OA *CanRisk Tool* from the University of Cambridge which is in the process of being commercialized. Its goal, in contrast to PENN II, is to compute the risk of breast cancer based on a variety of genetic and non-genetic indicators. That is, BOADICEA is concerned, more broadly than PENN II, with general cancer risk and not only HBOC syndrome related *BRCA1/2* mutation probabilities. Nonetheless, BOADICEA incorporates indicators based on family history and computes respective probabilities as a part of its analysis. Its questionnaire – still available for free at the moment after registration – contains questions about the following HBOC-related risk factors (cf. Figure 2): presence of BC, contralateral BC, OC and/or pancreatic cancer; family history including relation degrees and ages of cancer diagnosis; ancestry (again differentiating only between Ashkenazi-Jewish or non-Ashkenazi-Jewish). The questionnaire contains many more questions concerning other (non-genetic) factors. Note that although the ages can be specified as intervals, arithmetic means are used for further computations.



Medical History

Have you ever been diagnosed with endometriosis?

○ Yes ○ No ● Unknown

Have you had your tubes tied (tubal ligation)?

○ Yes ○ No ● Unknown

Have you ever had any of the following operations?	
Both ovaries removed (Oophorectomy)†	No
Both breasts removed (Mastectomy)	No

† Please note for women that have undergone BSO pre-menopausally, it can be assumed that their periods stopped at the age at BSO (in the Women's Health section). The current data on the effect of "natural" and "BSO-induced" menopause on breast cancer risk, suggest similar associations [Lancet Oncol. 2012].

Have you ever had any of these cancers?	
Breast cancer in one breast Please exclude Ductal Carcinoma-In-Situ (DCIS)	Yes Age: 35
Breast cancer in the second breast (contralateral breast cancer) each breast had its own cancer Please exclude Ductal Carcinoma-In-Situ (DCIS)	No
Ovarian cancer	No
Pancreatic cancer	No

Polygenic Risk Score(s)

Has a SNP array / Polygenic Risk Score (PRS), ever been run?

No

Family History

Have you any Ashkenazi Jewish ancestors?

○ Ashkenazi Jewish ● Other / Unknown

! HINT !
The **'Load'** button at the top of the tool can be used to load a pedigree file (e.g. BOADICEA file format exported from phenotips).

Individual's Details

Name: me

○ Male ● Female

● Alive ○ Deceased

Age: 35

Year of birth: 1988

Proband: ☑ Exclude: □

approx: ☑

Breast: 30-40

Contralateral breast: -

Ovarian: -

Pancreatic: -

Figure 2: Screenshot of BOADICEA web interface (a user account needed) from https://ccge.medschl.cam.ac.uk/boadicea/ showing only HBOC-related questions for the disease pattern considered in this subsection

At the moment, such general-purpose AI systems as ChatGPT[9] are not able to provide a differentiated answer about HBOC based on available publications. ChatGPT can be used to summarize and explain knowledge (with various challenges arising from this in scientific research and exchange), see [Schäfer 2023], therefore, it is in principle capable of finding out the appropriate information. For example, a broadly formulated question

[9] https://chat.openai.com/, May 3 version, accessed on May 8 2023

> Are you capable of giving predictions on genetic mutation risk?

is answered by ChatGPT as "As an AI language model, I am capable of providing information and explanations about genetic mutations and their associated risks based on available scientific data", with an additional caution to consult a gene specialist. Confronted with a concrete disease pattern, it answers mostly in general terms as

> "The presence of a family history of breast cancer can indicate an increased risk of developing the disease".

Only one (partial) pattern out of those we consider in Section 4 elicited actual probabilities as answers. The answer to the question

> What's the probability of carrying a pathogenic BRCA variant if a woman was diagnosed with breast cancer under the age of 40?

is

> ..."The exact probability of carrying a pathogenic BRCA1 or BRCA2 variant can vary depending on a variety of factors, including personal and family medical history, ethnicity, and other factors. However, some studies have estimated that the frequency of BRCA1 or BRCA2 mutations in individuals diagnosed with breast cancer before the age of 40 ranges from 8% to 17%",

again with a caution that "Genetic counseling and testing can provide more personalized information about an individual's specific risks and can inform decisions about management and prevention strategies". ChatGPT does not provide information about the studies it mentions. As a comparison, PENN II computes the individual risk of *BRCA1* mutation as 5%, *BRCA2* as 6% (equal to the familial risk in this case), BOADICEA as 3.19% and 2.93%, respectively. As we can see, the probabilities suggested by all of the tools are different, posing the question of which one to trust.

Obviously, the ChatGPT answers are quite impressive for a general-purpose AI system. Nonetheless, they still indicate the need for specific data mining (AI based) strategies in OA publications devoted to reporting *BRCA1/2* frequencies from the available literature (cf. Subsection 3.3, 3.4). They should be combined with mathematical models taking into account both aleatory and epistemic uncertainty in data for computing probabilities to be used during genetic counseling with a person or a family (cf. Section 4).

3 Design of an Optimal HBOC Database

The field of *BRCA1/2* research is developing rapidly. Recent technological advancements such as next generation sequencing (NGS) have enabled fast variant discovery in samples donated by cancer patients. Since not all variants necessarily cause cancer, they first need to be assessed by experts to classify their pathogenicity [Plon et al. 2008]. There are numerous databases dedicated to the classification of (*BRCA1/2*) variants available online: BRCA Exchange[10], ARUP *BRCA1/2* [11], LOVD[12], ClinVar[13]. These data can

[10] https://brcaexchange.org/

[11] https://arup.utah.edu/database/BRCA/

[12] https://www.lovd.nl/

[13] https://www.ncbi.nlm.nih.gov/clinvar/

help patients who already underwent genetic testing to understand their test results better. During genetic counseling, however, this information is not available yet. Therefore, risk assessment tools have to be able to estimate the mutation probability based only on knowledge about indicator factors in personal or family history.

While mutation risk can be estimated by relying on different theories (e.g., logistic regression, Bayesian networks, decision trees, cf. the overview in [Auer and Luther 2021]), the common feature of all strategies is the link between clinical data and variants' classification. All RA models usually need empirical data (ground truth). To obtain it, modeling is preceded by an extensive long-term survey in the best case. If this is not possible, the data from previous studies can be consolidated for the same purpose within a meta-study [Wang et al. 2021]. However, general standards about how to collect and report empirical data or record the meta-data about them are mostly missing, making definite conclusions about ground truth very difficult. As described in the introduction, the sparse availability of open-access cancer-related medical records for HBOC complicates the development of new RA models even further.

In this section, we first introduce our understanding of the concepts of data integration and data fusion (which are often used synonymously in literature but are actually not exactly the same) along with data provenance in Subsection 3.1. After that we propose a database design for storing ground truth in the context of the HBOC syndrome which can be understood as a first step towards a standard in Subsection 3.2. Note that the existing standards such as FHIR[14] or OMOP[15] cannot be used as they are, cf. [Bönisch et al. 2022]. The focus of the design proposed here is on flexibility: the model developers have to be able to perform the task of data fusion easily depending on the criteria they need the ground truth data on. Because *BRCA1/2* genes are the most researched ones in connection with breast and ovarian cancer, we consider only them, although the proposed database scheme can be extended to take into account further genes, for example, *CHEK2* or *PALB2*. In Subsection 3.3, we suggest an approach to extract information from OA publications into a database. This helps to avoid issues of classified information nature still present in the suggestion from 3.2. It is also a good way for making the data accessible without violating the patients' privacy. A proof of concept implementation illustrates the applicability of the data structure within the context of a DST based model for predicting *BRCA1/2* mutation probabilities.

3.1 Extracting data: Integration versus fusion; provenance

If several data sources are to be combined, the most important sub-processes to consider besides the data cleaning are

- consolidating various data schemata and
- combining the data objects.

If two or more SQL databases are considered as sources, then these sub-processes correspond roughly to the operations JOIN and MERGE, respectively. If, however, the data originate from different sources, the operations needed to be carried out are usually more complex, corresponding to *data integration* and *data fusion*. These terms are sometimes used interchangeably in the literature. However, they are separate, although

[14] https://fhir.org/

[15] https://www.ohdsi.org/data-standardization/

interconnected, processes, with data integration needing to be carried out prior to data fusion.

Data integration is extension of the structure and content of one data source by that of another [Li Lee and Ling 1995]. The basis of this operation is the so-called schema matching that checks the attributes of data sources for the three cases *identity*, *similarity*, *newness* and generates a global schema based on the findings using a top-down (e.g., by data harmonization) or a bottom-up approach depending on whether the number of data sources to combine is known beforehand [Arfaoui and Akaichi 2015].

The task of data integration is to build up a large body of information from many sources and to reduce it to a common data schema. By contrast, various data about one object have to be aggregated into a single data item during data fusion, for example, to combine multiple opinions on a particular issue into one statement. The data set to be combined is not allowed to contain duplicates, which might make further pre-processing necessary [Naumann et al. 2006, Dong et al. 2015]. If fusion is to be applied to non-numeric data, then the strategies of interaction (users decide which data is retained), selection (only the data from pre-selected sources is retained) and voting (data with the most frequent occurrence is retained) can be used. When working with numerical data, data synthesis using, for example, Bayesian networks, DST or fuzzy logic is to be preferred.

Not infrequently, the operations mentioned above are performed without giving much thought to the reliability of the data in the global database obtained in such a way. One part of making data reliable and helping to interpret data better is a record of their *provenance*. Similar approaches are being actively developed for intelligent log management of distributed applications (cf. [Harutyunyan et al. 2019]) and logging of data in general [Moreau et al. 2008] since many years.

Provenance is defined as “information about entities, activities, and people involved in producing a piece of data or thing, which can be used to form assessments about its quality, reliability or trustworthiness”[16]. Moreover, data provenance contributes to reproducibility of scientific results through “both systematic and formal records of the relationships among data sources, processes, datasets, publications and researchers” [Pasquier et al. 2017]. Provenance means that scientists require for their research not only the data themselves but also the meta-data about the data. A possible meta-data representation is through an acyclic directed graph with its vertices reflecting the involved persons/organizations, data items and data transformations and its edges corresponding to the interactions between them. This can be the basis for exploiting the provenance involving the necessary stages of meta-data capture, storage and analysis/visualization. Although such standards as PROV [Missier 2017] exist, they are still not used widely, especially while recording or reporting medical data. Standardizing medical science wrt. provenance guidelines is a challenging task, to tackle which is nonetheless important and necessary as pointed out and partially attempted, for example, in [Martínez-García and Hernandez-Lemus 2022] (without using the word ‘provenance’, exactly) and [Gierend et al. 2023].

3.2 Requirements and possible design of a traceable HBOC database

Over the past decades, researchers have agreed on certain factors that act as indicators for pathogenic *BRCA1/2* variants. Aside from a patient’s biological sex, their nationality and ethnicity can be decisive. As suggested in previous studies [Hall et al. 2009, Ashton-Prolla

[16] https://www.w3.org/TR/prov-overview/

and Vargas 2014], certain ethnic groups – most prominently people of Ashkenazi-Jewish ancestry – are at a higher risk of carrying *BRCA1/2* mutations, mostly due to endogamy and geographically contained gene pools (so-called "founder mutations"). Since it is possible that a person is a mutation carrier but has not developed cancer yet (while other family members have), the ideal HBOC database should be able to store relationships between individuals. Moreover, the age of cancer patients at the time of diagnosis, the primary tumor location and subtype, as well as all variants occurring in *BRCA1/2* genes must be recorded. Registering every variant is necessary because its classification as pathogenic or non-pathogenic can change over time. It should also be indicated if a somatic or germline mutation occurred.

There are existing standards for recording and exchanging medical information such as FHIR or OMOP. However, the common goal of these standards is to keep a reliable record, for one actual patient, of the development of their disease history with time, including details on observations, conditions, prescriptions in order to provide best care. These are not the models on which databases like GDC or BRCA Exchange, relevant for the research in this paper, are based[17]. Such details are not necessary for obtaining information on mutation probabilities and would additionally disclose information openly violating data privacy legislation which is what we try to avoid as far as possible in our suggestion for a traceable database. Moreover, it has been demonstrated recently in [Bönisch et al. 2022] that "none of the data formats include all metadata, which is required to successfully operate the MeDIC[18] for the purpose of reliable data management". Hence, we formulate the requirements specific to the HBOC research which can be understood as a first step towards a standard in this area. Obviously, there is a need for automated compression and anonymization of FHIR or OMOP patient data into any data format suitable for research on HBOC.

In [Auer and Luther 2022], we came to the conclusion based on the extensive analysis of the available publications that, in order to be able to perform data integration or fusion, it is necessary to choose studies with clearly described cohorts of large sizes which classify the included patients and their family members wrt.

– the risk factor single/multiple breast cancer (also, in the same person), bilateral, male breast cancer and ovarian cancer with the record about the respective first age of onset;
– standardized selection criteria (disease patterns), the a priori risk class;
– origin/ethnicity of the patient (e.g., the youngest family member with BC), family history including the degree of relationship and the first occurrence of the disease;
– detected tumor subtypes (e.g., triple negative breast cancer); and, finally,
– the quality and the trustworthiness of the data (e.g., use of public databases with a disclosed search strategy), ideally, its provenance or lineage.

However, the list of the important risk factors or selection criteria might change with the new developments in medical research. A future standard should be flexible enough to incorporate any such changes.

We consider a relational database design consisting of eight tables to be suitable for the task at hand: four tables for the entities `cancer`, `person`, `variant` and `project`, and four relational tables for the connections between them. The corresponding entity-relationship diagram is shown in Figure 3. This architecture allows multiple cases of

[17] See https://gdc.cancer.gov/developers/gdc-data-model, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6324924/

[18] Medical Data Integration Center

cancer in the same person, while a set of variants can be assigned to each patient, describing only the necessary genetic properties in the context of *BRCA1/2* genes.

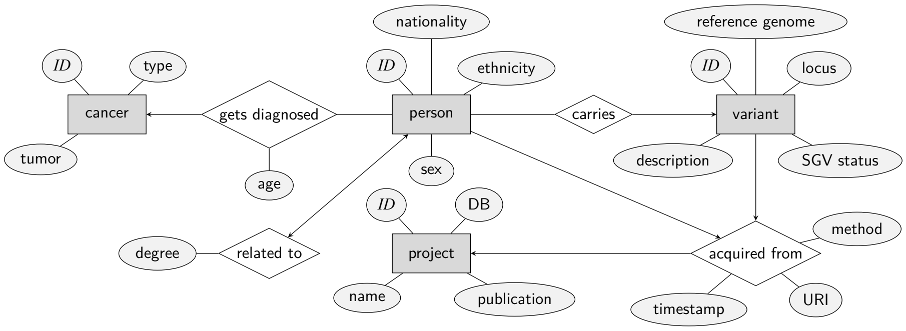


Figure 3: Entity-relationship diagram

By combining oncogenic databases with variant databases using data integration strategies, the standard HBOC database could be constructed without the need for conducting new large-scale, long-term surveys while retaining a good degree of data provenance because all data fusion instances possibly carried out for obtaining concrete values for risk factors can be traced to the corresponding database items as well as aggregation strategy be made clear or exchanged. As shown in Figure 3, each personal record should be associated with a project, to trace back its original study cohort. In well-established oncogenic databases like GDC and ICGC it is common practice to disclose research projects' goals and methodologies to ensure a record's credibility. Similarly, it must be transparent how a variant was discovered and classified. Patient records and variants both originate from third-party sources, so it is essential to document time and method of acquisition, as well as a unique resource identifier (URI) for online databases. The umbrella term `project` is used here for the research project that published a data record in a public database and possibly summarized its findings in an accompanying publication referred to by its DOI.

Combining the two types of databases considered in this scenario can be expressed as a two-way extract-transform-load (ETL) process shown in Figure 4. The ETL process is standard in data integration/warehousing tasks [Denney et al. 2016] and consists of the following three steps:

First, relevant subsets are queried from all data sources. Although containing common domain-specific data, different databases often use distinct technology stacks, which may require a query be translated. Such a 'translation' restructures a single request to meet the grammatical rules of various other query languages. In our context, this step applies to both oncogenic and variant databases and is needed to extract all data on both *BRCA1/2* -related cancers and variants, respectively.

Second, the transformation step ensures that all extracted records follow the same global schema. This can be achieved by either schema matching (finding similar data fields and rearranging them in a common structure) or schema mapping (fitting similar data fields into a predefined data model). In our context, schema mapping is preferred,

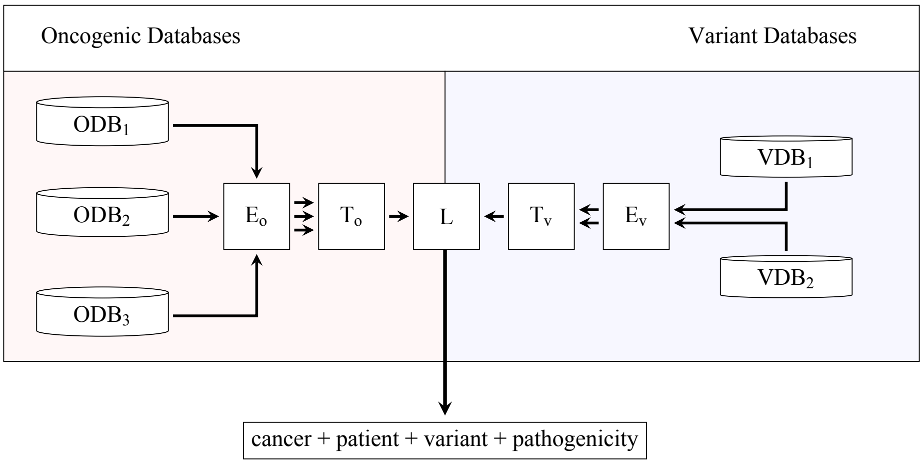


Figure 4: Two-way ETL scheme

since the required tables and fields are already known (cf. Figure 3). Up to this point, oncogenic and variant databases are handled separately and produce two homogeneous integrated datasets.

Finally, oncogenic and variant data need to be joined, based on variant names. This part is probably rather complicated because of synonymous nomenclature used for a single variant across multiple sources; possible duplicates would have to be eliminated. Note that the database BRCA Exchange actually stores synonyms for each registered variant and might be used as a thesaurus in this case. Once oncogenic patient records, variants and their pathogenicity have been linked, the merged dataset can be loaded thus providing absolute mutation frequencies as the basis for a new risk assessment model.

Although this design stores only a minimum amount of personal data, the question of privacy still arises because of the short DNA excerpts linked to patients, which may again require that access be restricted. In the next subsection we suggest an alternative database possibility to circumvent any privacy concerns.

3.3 Construction of an alternative database from scientific literature

The main issue with open access genetic data is the possibility of personal identification, even though data sources might be used only to calculate relative frequencies of mutation carriers among a group of patients with similar characteristics. The mentioned frequencies themselves, however, may in fact be published, because they do not contain any sensitive data anymore. To make use of this fact, we propose a second database design that does not depend on the first-hand genetic data at all. Instead, we utilize findings from OA publications by scientists with access to genetic samples linked to medical information and build a database upon findings from multiple sources.

Note that the alternative suggested here is a proposal of how to publish classified data without privacy violation concerns, achieved through aggregation. Although the degree of trustworthiness diminishes, this would give those developers of new RA models who are not linked to big scale medical institutions at least some semblance of traceable data. The proposal from Subsection 3.2 is to be preferred if researchers are granted access to sensitive data. The approach from this subsection is an acceptable compromise if not. It provides data in a standardized form and documents meta-data to a certain degree by recording all source publications and subsequent aggregation strategies.

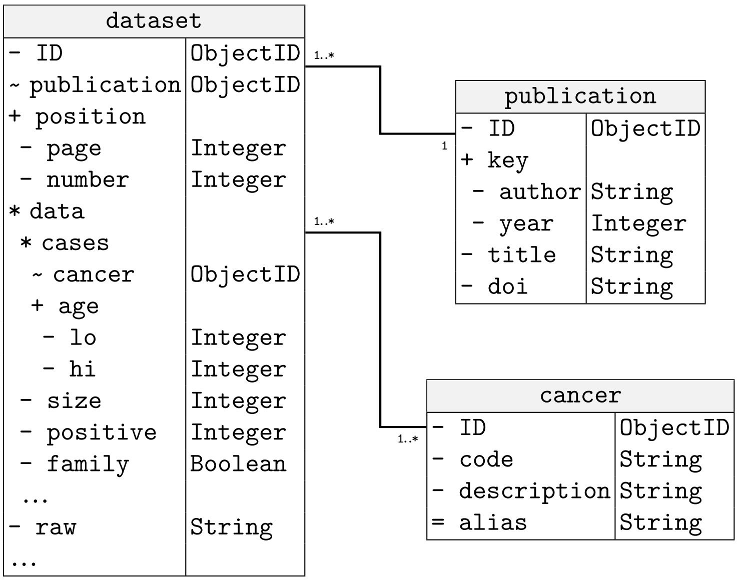
While developing the database, we focused primarily on data in the form of tables from medical publications. Obviously, graphical representations such as histograms contain valuable information, too, but are much more difficult to translate into exact data, both manually and automatically. Usually, the prevalence of *BRCA1/2* mutations presented in such research papers is stratified by attributes similar to those described at the beginning of Subsection 3.2 and is given as absolute frequencies. Because various publications might focus on different types of cancer, patient's characteristics or age intervals, the database used for storing the extracted information must support heterogenous data structures. Therefore, we recommend the use of a NoSQL database. The data extracted from one publication may be structured in the following way:

```
[
  {
    case: {
      cancer: 'breast',
      type: 'bilateral',
      age: { lo: 18, hi: 29 }
    },
    total: 26236,
    genes: [
      { brca1: 4996 },
      { brca2: 2519 }
    ]
  }
]
```

The flexible nature of NoSQL databases makes using traditional database design principles to describe their entire structure somewhat difficult. Since records in a NoSQL database should be considered as linked/related objects of similar appearance rather than rows in a strictly constrained table, object-oriented modeling paradigms can be applied. The modeling of NoSQL databases is still a subject of ongoing research, with no established standard as yet. In Figure 5, we use an extension of the UML class diagram standard [Sparks 2011] to represent the database structure because of its widespread familiarity and readability. Its use of symbols has been modified to express an object's structure in the following way:

−	atomic property; attribute in the classical sense
=	multiple instances of the same atomic property
+	composite property; the attribute itself is an object
∗	enumeration of composite properties
~	reference to another object
…	anything.

Note that object nesting, a common operation in object-oriented databases, can also be represented as a classical relational model. However, such kind of a representation can be misinterpreted as a system of linked objects, which is not the case in our database. Objects of nested attributes can still have relations, while nested attributes are usually entirely private to the parent objects' scope.



dataset	
- ID	ObjectID
~ publication	ObjectID
+ position	
- page	Integer
- number	Integer
* data	
* cases	
~ cancer	ObjectID
+ age	
- lo	Integer
- hi	Integer
- size	Integer
- positive	Integer
- family	Boolean
...	
- raw	String
...	

publication	
- ID	ObjectID
+ key	
- author	String
- year	Integer
- title	String
- doi	String

cancer	
- ID	ObjectID
- code	String
- description	String
= alias	String

Figure 5: Publication database schema

The proposed database architecture in Figure 5 links tabular data (`datasets`) to its corresponding `publication` object, which in turn references the original source material by its DOI to keep the data traceable. Every object of type `dataset` is a single table with a unique position (`page` being the relative page number in the linked publication; `number` refers to the absolute table count per page). Contextually relevant table cells are described as `data` objects, covering the `positive` rate of a certain type of `cancer` (with optional `aliases` to compensate for synonyms) for a cohort of sample size `size` in the `age` bracket `lo` to `hi`, and recording whether or not this sample concerns the `family` history. The original `raw` data should be present for validation.

3.4 Proof of concept literature database for computing *BRCA1/2* mutation probabilities

We tested the database design proposed in Subsection 3.3 wrt. its practicability by implementing it in the NoSQL database system MongoDB. Its extensive set of aggregation functions allowed us to perform most computations inside the database itself, for example, for construction of age intervals for certain risk factors or for retrieval of the relative frequency of pathogenic variants in every calculated interval. The data fusion of multiple sources for the same risk factors can be used to produce a general approximation of the mutation probability depending on different criteria.

Similar to [Auer and Luther 2021], the general mutation risk was modeled as a combined Dempster-Shafer structure of the proband's personal history and their family history. Partitioned into those two categories, each data source was weighted based on additional quality criteria. A source was considered more important if its sample size was sufficiently large with clearly divided age groups and a transparent classification strategy. For each case of cancer in a specific age group, its mass was set to the weighted average (cf. Eq. (4)) across all relevant sources, for the personal (m_p) and familial (m_f) mutation risk respectively. Familial and personal mass distributions were combined into a single Dempster-Shafer structure (m_{pf}), using Dempster's rule (cf. Eq. (3)). We used the R language to combine the aggregated mass assignments, and the `mongolite` package as an interface to our database. In Table 1, the obtained BPAs are displayed. A screenshot of the web application is shown in Figure 6.

Case	m_p	m_f	m_{pf}
$\text{BC}_{<40}$	0.020	0.043	0.006
$\text{BC}_{<50}$	0.075	0.030	0.015
$\text{BC}_{\geq 50}$	0.034	0.108	0.024
$\text{BC}_{\text{bilateral}}$	0.030	0.088	0.017
BC_{male}	0.037	0.138	0.033
OC	0.151	0.104	0.103
$\text{BC}_{<50}$, OC	0.058	0.054	0.020
$\text{BC}_{\geq 50}$, OC	0.119	0.245	0.191
Ω	0.476	0.190	0.591

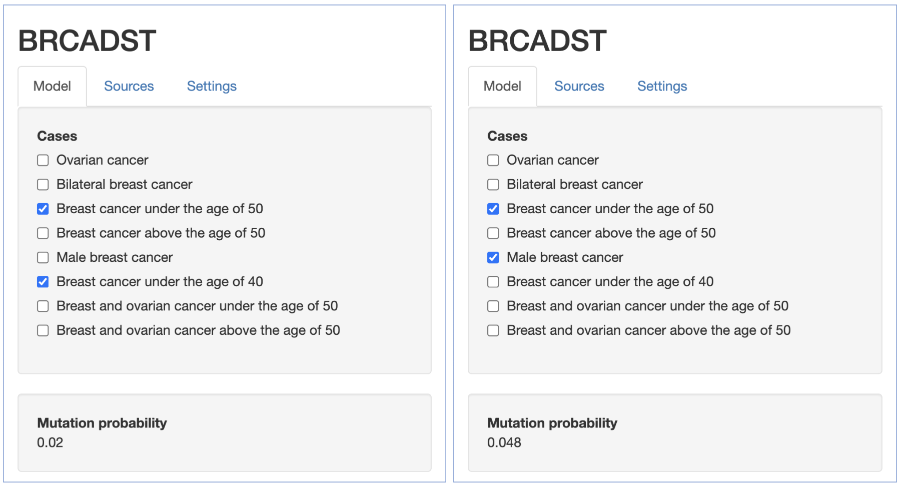
Table 1: Aggregated BPAs from publications

Due to limited data availability, only few scenarios were covered by this model (cf. Table 3). Additionally, poor source data quality led to lower weights in some cases, which in turn reduced the mutation risk significantly in comparison with the existing models. For example, the risk of a person whose father was diagnosed with breast cancer before the age of 50 would be calculated as $Bel_{m_{pf}}(\{\text{BC}_{<50}, \text{BC}_{\text{male}}\}) = m_{pf}(\text{BC}_{<50}) + m_{pf}(\text{BC}_{\text{male}}) =$ 4.8%, as compared to 10 – 18% predicted by PENN II. That is, better data are certainly needed but the frequencies can definitely be obtained as proposed in Subsection 3.3.

Our limited test data allowed a differentiation per case only between "below the age of 50" and "above the age of 50" at best. Since the age of onset [Buys et al. 2017] strongly influences the likelihood of being a carrier of a pathogenic *BRCA1/2* variant, we decided that a more granular approach to handling the overall risk as a function in dependence of a patient's age was needed. A model for this is proposed in the following section.

In this first implementation, data acquisition for testing was performed manually. We found that there was still a need for more sophisticated AI-based text mining tools if the process is to be automated. While the problem of table extraction from text (PDF files in our case) has been solved[19], automatic interpretation of extracted data in a certain context is a challenge yet to be overcome and a topic for future research.

[19] cf. https://cran.r-project.org/web/packages/PDE/vignettes/PDE.html



BRCADST

Cases

- ☐ Ovarian cancer
- ☐ Bilateral breast cancer
- ☑ Breast cancer under the age of 50
- ☐ Breast cancer above the age of 50
- ☐ Male breast cancer
- ☑ Breast cancer under the age of 40
- ☐ Breast and ovarian cancer under the age of 50
- ☐ Breast and ovarian cancer above the age of 50

Mutation probability
0.02

BRCADST

Cases

- ☐ Ovarian cancer
- ☐ Bilateral breast cancer
- ☑ Breast cancer under the age of 50
- ☐ Breast cancer above the age of 50
- ☑ Male breast cancer
- ☐ Breast cancer under the age of 40
- ☐ Breast and ovarian cancer under the age of 50
- ☐ Breast and ovarian cancer above the age of 50

Mutation probability
0.048

Figure 6: Screenshots of the web app based on the data structure from Subsection 3.3 for the disease pattern considered in Subsection 2.3 (left) and a further pattern from this subsection

4 Modeling the *BRCA1/2* Mutation Probabilities as a Function of Age

As mentioned above, it is necessary to differentiate better in dependence on age in the developed DST model to obtain more realistic results. Since we were not able to arrive at ground truth for HBOC risk factors from the databases available on the internet due to privacy restrictions (cf. Subsection 3.2), we use the manually collected information from OA publications [Frank et al. 2002, Buys et al. 2017], supplemented by predictions from PENN II where necessary, directly. In this section, we present a proof of concept DST approach to assess the probability of *BRCA1/2* mutation based on finer age models. We consider it to be merely a proof of concept since the data we rely on are selective and possibly do not reflect ground truth yet. Nonetheless, we are able to achieve good correspondence with the state of the art systems, which are based on (partially) undisclosed data (cf. Subsection 2.3). Note that better data can be directly incorporated into the proposed model.

The age at the first cancer diagnosis is one of the most important indicators for the presence of a *BRCA1/2* mutation [Buys et al. 2017]. As suggested in [Auer and Luther 2022], the mutation probability can be modeled by combining age-based cumulative percentage curves for different risk factors. In this paper, we extend this idea by utilizing the Dempster-Shafer theory of evidence as the basis for combining multiple risk factors in dependence on the patient's age and the age of his/her relatives with a history of *BRCA1/2* -related cancer.

We consider the risk factors BC, male breast cancer (mBC), bilateral breast cancer (bBC), OC, breast and ovarian cancer (BCOC) in the history, BC and OC cancer in the same person (BCOCsp), and ethnicity (E) for ages between 60 and 20 in steps of 5 years. At the moment, we only differentiate between two ethnicities: general and Ashkenazi-Jewish (AJ). The masses for BPAs at each given age are shown in Table 2. Note the curves for risk indicators BCOC and BCOCsp are obtained by considering the sum of values for BC and OC and the additive factor sp, respectively. For all ages in the case of Ashkenazi-Jewish ethnicity, additive factors of 0.05 are used for BC and bBC; of 0.03 for OC; and of 0.01 for sp. For obtaining the cumulative probability curves shown in Figure 7 (general on the left, AJ on the right), we interpolate linearly between the values for the ages (as is usually done in statistics). In the same way as in the implementation from Subsection 3.4, we assume that the focal element Ω containing all the considered risk factors is assigned the remainder of the probability since we compute only the lower bound on the risk (which, however, can itself be an interval, cf. Subsection 2.2). All other subsets of Ω aside from those in Table 2 (and Ω) are supposed to have zero masses.

Age	**BC**	**bBC**	**mBC**	**OC**	**sp**
60	0.04	0.02	0.09	0.03	0.03
55	0.015	0.02	0.09	0.05	0.02
50	0.015	0.02	0.09	0.07	0.02
45	0.02	0.02	0.09	0.09	0.02
40	0.02	0.02	0.09	0.11	0.02
35	0.02	0.02	0.09	0.13	0.02
30	0.03	0.02	0.09	0.15	0.02
25	0.04	0.02	0.09	0.17	0.02
20	0.04	0.02	0.09	0.19	0.02

Table 2: BPAs for the chosen risk factors wrt. age

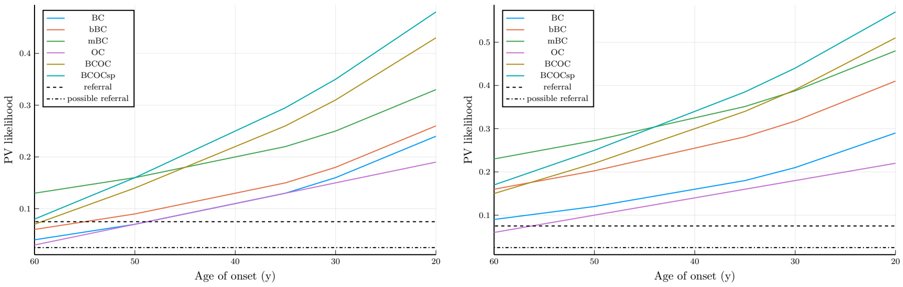


Figure 7: Cumulative risk curves for the mutation probabilities: general (left) and AJ (right) case

In Figure 7, there are two lines showing the threshold values for considered and recommended gene test referral (at 2.5% and 7.5%, respectively) that are given by a relatively recent meta-study [Pujol et al. 2021]. As the cumulative curves show, practically all individual mutation probabilities (e.g., as reported in [Buys et al. 2017]) are higher than those thresholds. In our opinion, it is necessary to reconsider those values, at least, if such models as PENN II are to be viewed as trustworthy.

Our model works as follows: For each examined pattern, one BPA is usually constructed for the individual (m_p) and a further one for the person in her family history (m_h) based on the cumulative curves in dependence on the ages given by a disease pattern and the risk indicator (m_i for $i \in \{$BC, bBC, mBC, OC, sp$\}$). In cases where the age of onset for a family member is not known exactly, we use intervals for the probabilities (IBPAs). Because of the monotonic properties of the cumulative risk curves, we can assume the interval interpretation of the mass function $m_i(.)$ for a risk factor i to be $[m_i(\overline{a}), m_i(\underline{a})]$, $[\underline{a}, \overline{a}]$ being the age interval. If there are more than one individual mentioned in the family history, it is possible to consider additional BPAs for each mentioned person. At the moment, all (I)BPAs have the same importance and are merged using Dempster's rule of combination (cf. Eq. (3)). However, we plan to implement the possibility to apply Mendelian logic for aggregating since there is no difficulty to realize it in our model in principle (e.g., using the mixing and averaging rule, cf. Eq. (4)). In the final step, the belief function is calculated as given by Eq. (2), on the right, for all applicable risk factors employing the combined mass assignment m_{pf}.

We apply our model ("new") to the example disease patterns described below and compare them with the results from [Auer and Luther 2022] ("old"), with the app described in Subsection 3.4 ("app") where possible and with those from PENN II, BOADICEA and Frank tables where available (cf. Table 3). The shown intervals are rounded outwards to three decimal positions; crisp numbers are rounded to three decimal positions using rounding to the next number where necessary.

Non-Ashkenazi-Jewish

Pattern 1: Father with BC at 42, second degree relative with BC>50

Pattern 2: Patient with BC<40 and mother with BC<50

Pattern 3: Patient with OC<40 and her aunt with BC<50

Pattern 4: Patient aged 22 with BC and OC; mother with bBC>50

Ashkenazi-Jewish

Pattern 5: Patient (30-40 years of age) with BC, aunt with BC>51

Pattern 6: Patient with BC at 45 years of age, sister with OC<50

Pattern 7: Patient with OC<50 and OC and BC>50 in the family history

Pattern 8: Patient with OC and BC at 35 years of age, aunt with bBC over 50

The results for the new model show a good agreement with those from [Frank et al. 2002] and PENN II, which is not surprising: Although the models are different, we still rely on and incorporate data from those two sources into our model as ground truth. The results from BOADICEA show a good agreement for Patterns 4 and 5 only (they are also quite different from PENN II). Note that BOADICEA is designed to assess the general risk of contracting breast/ovarian cancer from a variety of genetic and non-genetic indicators, among others due to hereditary factors. That means it is not easy to reflect the considered disease patterns in exactly the same way as described above since a lot more questions about the patient need to be answered. Besides, both PENN II and BOADICEA are very sensitive wrt. the age of onset for the cancer cases (making the need of interval representations even more important). That is, the values for PENN II and BOADICEA

Pa.	New	Old	PENN II	BOADICEA	Frank	App
1	[0.251, 0.294]	0.257	0.27	0.059		0.072
2	[0.116, 0.317]	0.198	0.15 – 0.35	0.073	0.297	0.044
3	[0.232, 0.584]	0.358	0.12 – 0.4	0.036		0.137
4	[0.446, 0.549]	0.527	0.54	0.491		0.184
5	[0.149, 0.275]	0.319	0.24 – 0.35	0.298	0.318	
6	[0.296, 0.406]	0.313	0.31	0.662	0.415	
7	[0.225, 0.410]	0.417	0.41	0.727	0.412	
8	[0.459, 0.603]	[0.518,0.656]	0.53	0.937		

Table 3: BRCA1/2 mutation probabilities for the example disease patterns

from Table 3 are provided as a background reference and reflect the results that might be obtained by a non-expert patient filling out the respective questionnaires; they might change if a gene expert operates the interfaces. The results from the app (last column) show poor agreement systematically underestimating the risk for reasons explained in Subsection 3.4, although they are based on the same model as in [Auer and Luther 2022] (Column "old" with a medium to good agreement), demonstrating the importance of good quality data on ground truth.

The advantage of the new model proposed in this paper is that it now also supplies the lower (upper) bound on the belief function and not just an average or median value. Besides, since it can be made to work with the database from Subsection 3.3, it can incorporate traceable data. In this way, decisions about the computed probabilities can be explained comprehensively.

5 Conclusions and Future Work

In this paper, we made a first step towards a standard for collecting and storing HBOC-related data with a focus on its flexibility and reliability through provenance as well as methods with result verification. In Subsections 3.2 and 3.3, we suggested two database designs for data on HBOC, one containing classified information and one avoiding the restrictions due to privacy concerns through aggregation, respectively. The data extraction approach from 3.2 was designed but not yet implemented for HBOC, exactly because of privacy legislation restrictions. In Subsection 3.4, we implemented the design from 3.3 to incorporate data from several OA publications. We showed that the proposed database structure was well suited for obtaining aggregated frequencies for risk indicators needed, for example, in the context of a DST model for computing *BRCA1/2* mutation probabilities first introduced in [Auer and Luther 2021]. Lessons learned from working with this model were that a more graded approach to modeling risk indicators in dependence on the age of first cancer occurrence was necessary, resulting in the two-stage DST based technique presented in Section 4. It takes into account epistemic uncertainty by considering intervals if ages of persons from the family history are not known exactly and employs IA with result verification for all involved (DST-related) operations.

Although this implementation shows a good agreement with established tools while providing more information on *BRCA1/2* mutation probabilities through working with interval BPAs, there is still room for improvement. For example, this approach does not take into account the degree of relationship, that is, individual and family history are treated as equally relevant at the moment. Future improvements might include possible redistribution using further combination rules (aside from Dempster's rule) to account for inheritance probability. Another interesting topic for future research is to apply interval arithmetic type 2 in the context of this DST approach and compare with the 'normal' IA version wrt. both complexity and prediction accuracy.

Yet another point is that the issue of finding ground truth about HBOC syndrome from OA sources automatically is still largely unresolved due to multiple reasons starting with privacy issues and ending with the lack of related standards. As a manageable topic for future research, we will study the possibility of employing customized AI for data extraction about HBOC from OA papers.

References

[Arfaoui and Akaichi 2015] Arfaoui, N., Akaichi, J.: "Automating Schema Integration Technique Case Study: Generating Data Warehouse Schema from Data Mart Schemas"; Kozielski, S., Mrozek, D., Kasprowski, P., Malysiak-Mrozek, B., Kostrzewa, D. (eds): Proceedings of 11^{th} International Conference: Beyond Databases, Architectures and Structures, Communications in Computer and Information Science, 521, Springer, 2015, 200–209.

[Ashton-Prolla and Vargas 2014] Ashton-Prolla, P., Vargas, F.: "Prevalence and impact of founder mutations in hereditary breast cancer in Latin America"; Genet. Mol. Biol., 37(1), 2014, 234–240.

[Auer and Luther 2021] Auer, E., Luther, W.: "Uncertainty Handling in Genetic Risk Assessment and Counseling"; JUCS - Journal of Universal Computer Science, 27(12), 2021, 1347–1370.

[Auer and Luther 2022] Auer, E., Luther, W.: "Dempster-Shafer Theory Based Uncertainty Models for Assessing Hereditary, BRCA1/2-Related Cancer Risk"; The 8^{th} International Symposium on Reliability Engineering and Risk Management, 2022, 755–762.

[Auer et al. 2010] Auer, E., Luther, W., Rebner, G., Limbourg, Ph.: "A Verified MATLAB Toolbox for the Dempster-Shafer Theory"; Proc. of the Workshop on the Theory of Belief Functions, 2010.

[Ayyub and Klir 2006] Ayyub, B., Klir, G.: Uncertainty Modeling and Analysis in Engineering and the Sciences; 2006.

[Bönisch et al. 2022] Bönisch, C., Kesztyüs, D., Kesztyüs, T.: "Harvesting metadata in clinical care: a crosswalk between FHIR, OMOP, CDISC and openEHR metadata"; Scientific Data, 2022, 659.

[Bradley 2019] Bradley, S.: "Imprecise Probabilities"; Zalta, E. N. (ed): The Stanford Encyclopedia of Philosophy, Spring 2019 edn, Metaphysics Research Lab, Stanford University.

[Buys et al. 2017] Buys, S. S., Sandbach, J. F., Gammon, A., Patel, G., Kidd, J., Brown, K. L., Sharma, L., Saam, J., Lancaster, J., Daly, M. B: "A study of over 35,000 women with breast cancer tested with a 25-gene panel of hereditary cancer genes"; Cancer, 123(10), 2017, 1721–1730.

[Claus et al. 1994] Claus, E. B., Risch, N., Thompson, W. D.: "Autosomal dominant inheritance of early-onset breast cancer. Implications for risk prediction"; Cancer, 73(3), 1994, 643–651.

[de Figueiredo and Stolfi 2004] de Figueiredo, L. H., Stolfi, J.: "Affine Arithmetic: Concepts and Applications"; Numerical Algorithms, 34(1–4), 2004, 147–158.

[Denney et al. 2016] Denney, M., Long, D., Armistead, M., Anderson, J., Conway, B.: "Validating the Extract, Transform, Load Process Used to Populate a Large Clinical Research Database"; International Journal of Medical Informatics, 94(07), 2016.

[Desrochers and Jaulin 2017] Desrochers, B., Jaulin, L.; "Thick Set Inversion"; Artificial Intelligence, 249, 2017, 1–18.

[Dong et al. 2015] Dong, X. L., Berti-Équille, L., Srivastava, D.: "Data Fusion: Resolving Conflicts from Multiple Sources"; CoRR, 2015.

[Ferson et al. 2003] Ferson, S., Kreinovich, V., Ginzburg, L., Myers, D. S., Sentz, K.: Constructing Probability Boxes and Dempster-Shafer Structures; Washington, D.C: United States. Dept. of Energy, 2003.

[Frank et al. 2002] Frank, T. S., Deffenbaugh, A. M., Reid, J. E., Hulick, M., Ward, B. E., Lingenfelter, B., Gumpper, K. L., Scholl, T., Tavtigian, S. V., Pruss, D. R., Critchfield, G. C.: "Clinical Characteristics of Individuals With Germline Mutations in BRCA1 and BRCA2: Analysis of 10,000 Individuals"; J. Clin. Oncol., 20(6), 2002, 1480–1490.

[Gierend et al. 2023] Gierend, K., Wodke, J. A.H., Genehr, S., Gött, R., Henkel, R., Krüger, F., Mandalka, M., Michaelis, L., Scheuerlein, A., Schröder, M., Zeleke, A., Waltemath, D.: "TAPP: Defining standard provenance information for clinical research data and workflows – Obstacles and opportunities"; Companion Proceedings of the ACM Web Conference 2023, ACM, 2023.

[Guerrini et al. 2017] Guerrini, C. J., McGuire, A. L., Majumder, M. A.: "Myriad take two: Can genomic databases remain secret?"; Science, 356(6338), 2017, 586–587.

[Hall et al. 2009] Hall, M. J., Reid, J. E., Burbidge, L. A., Pruss, D., Deffenbaugh, A. M., Frye, C., Wenstrup, R. J., Ward, B. E., Scholl, Th. A., Noll, W. W.: "*BRCA1* and *BRCA2* mutations in women of different ethnicities undergoing testing for hereditary breast-ovarian cancer"; Cancer, 115(10), 2009, 2222–2233.

[Harutyunyan et al. 2019] Harutyunyan, A. N., V. Poghosyan, A., M. Grigoryan, N., A. Hovhannisyan, N., Kushmerick, N.: "On Machine Learning Approaches for Automated Log Management"; JUCS - Journal of Universal Computer Science, 25(8), 2019, 925–945.

[Kahan 1968] Kahan, W.: A more complete interval arithmetic. Lecture notes for a summer course, University of Toronto, 1968, Canada.

[Lee et al. 2019] Lee, A., Mavaddat, N., Wilcox, A. N., Cunningham, A. P., Carver, T., Hartley, S., de Villiers, C. B., Izquierdo, A., Simard, J., Schmidt, M. K., Walter, F. M., Chatterjee, N., Garcia-Closas, M., Tischkowitz, M., Pharoah, P., Easton, D. F., Antoniou, A. C.: "BOADICEA: a comprehensive breast cancer risk prediction model incorporating genetic and nongenetic risk factors"; Genetics in Medicine, 21(8), 2019, 1708–1718.

[Li Lee and Ling 1995] Li Lee, M., Ling, T. W.: "Resolving structural conflicts in the integration of Entity-Relationship schemas"; Papazoglou, Michael P. (ed), OOER '95: Object-Oriented and Entity-Relationship Modeling. Berlin, Heidelberg: Springer Berlin Heidelberg, 1995, 424–433.

[Lindor et al. 2010] Lindor, N. M., Johnson, K. J., Harvey, H., Shane Pankratz, V., Domchek, S. M., Hunt, K., Wilson, M., Cathie Smith, M., Couch, F.: "Predicting BRCA1 and BRCA2 gene mutation carriers: comparison of PENN II model to previous study"; Familial Cancer, **9**(4), 2010, 495–502.

[Lohner 2001] Lohner, R.: "On the Ubiquity of the Wrapping Effect in the Computation of Error Bounds". Kulisch, U., Lohner, R., Facius, A. (eds), Perspectives on Enclosure Methods. Springer-Verlag, 2001, 201–218.

[Makino and Berz 2004] Makino, K., Berz, M.: Suppression of the Wrapping Effect by Taylor Model-Based Validated Integrators; MSUHEP 40910. Department of Physics, Michigan State University, East Lansing, MI 48824, 2004.

[Martínez-García and Hernandez-Lemus 2022] Martínez-García, M., Hernandez-Lemus, E.: "Data Integration Challenges for Machine Learning in Precision Medicine"; Frontiers in Medicine, 8(1), 2022.

[Merheb et al. 2013] Merheb, R., Mora, L., Palomo del Barrio, E.: "Parameter estimation in an uncertain and noisy environment via set inversion"; IEEE Symposium on Computational Intelligence for Engineering Solutions (CIES), 2013, 125-132.

[Missier 2017] Missier, P.: "Provenance Standards"; Liu, L., Özsu, M. T. (eds), Encyclopedia of Database Systems, Springer New York, 2017, 1–8.

[Moore et al. 2009] Moore, R. E., Kearfott, R. B., Cloud, M. J.: Introduction to Interval Analysis; Philadelphia: Society for Industrial and Applied Mathematics, 2009.

[Moreau et al. 2008] Moreau, L., Ludäscher, B., Altintas, I., Barga, R. S., Bowers, S., Callahan, S., Chin, Jr., G., Clifford, B., Cohen, S., Cohen-Boulakia, S., Davidson, S., Deelman, E., Digiampietri, L., Foster, I., Freire, J., Frew, J., Futrelle, J., Gibson, T., Gil, Y., Goble, C., Golbeck, J., Groth, P., Holland, D. A., Jiang, S., Kim, J., Koop, D., Krenek, A., McPhillips, T., Mehta, G., Miles, S., Metzger, D., Munroe, S., Myers, J., Plale, B., Podhorszki, N., Ratnakar, V., Santos, E., Scheidegger, C., Schuchardt, K., Seltzer, M., Simmhan, Y. L., Silva, C., Slaughter, P., Stephan, E., Stevens, R., Turi, D., Vo, H., Wilde, M., Zhao, J., Zhao, Y.: "Special Issue: The First Provenance Challenge"; Concurr. Comput. : Pract. Exper., 20(5), 2008, 409–418.

[Naumann et al. 2006] Naumann, F., Bilke, A., Bleiholder, J., Weis, M.: "Data Fusion in Three Steps: Resolving Schema, Tuple, and Value Inconsistencies"; IEEE Data Eng. Bull., 29(2), 2006 21–31.

[Neumaier 2003] Neumaier, A.: "Taylor Forms – Use and Limits"; Reliable Computing, 9(1), 2003, 43–79.

[Pasquier et al. 2017] Pasquier, T., Lau, M. K., Trisovic, A., Boose, E. R., Couturier, B., Crosas, M., Ellison, A. M., Gibson, V., Jones, C. R., Seltzer, M.: "If these data could talk"; Scientific Data, 4(1), 2017, 170114.

[Piegat and Dobryakova 2020] Piegat, A., Dobryakova, L.: "A Decomposition Approach to Type 2 Interval Arithmetic"; International Journal of Applied Mathematics and Computer Science, 30(03), 2020, 185–201.

[Plon et al. 2008] Plon, S. E., Eccles, D. M., Easton, D., Foulkes, W. D., Genuardi, M., Greenblatt, M. S., Hogervorst, F. B. L., Hoogerbrugge, N., Spurdle, A. B., Tavtigian, S. V., et al.: "Sequence variant classification and reporting: recommendations for improving the interpretation of cancer susceptibility genetic test results"; Human mutation, 29(11), 2008, 1282–1291.

[Pujol et al. 2021] Pujol, P., Barberis, M., Beer, P., Friedman, E., Piulats, J. M., Capoluongo, E.D., Garcia Foncillas, J., Ray-Coquard, I., Penault-Llorca, F., Foulkes, W. D., Turnbull, C., Hanson, H., Narod, S., Arun, B. K., Aapro, M. S., Mandel, J.-L., Normanno, N., Lambrechts, D., Vergote, I., Anahory, M., Baertschi, B., Baudry, K., Bignon, Y.-J., Bollet, M., Corsini, C., Cussenot, O., De la Motte Rouge, T., M.Duboys de Labarre, Duchamp, F., Duriez, C., Fizazi, K., Galibert, V., Gladieff, L., Gligorov, J., Hammel, P., Imbert-Bouteille, M., Jacot, W., Kogut-Kubiak, T., Lamy, P.-J., Nambot, S., Neuzillet, Y., Olschwang, S., Rebillard, X., Rey, J.-M., Rideau, C., Spano, J.-P., Thomas, F., Treilleux, I., Vandromme, M., Vendrell, J., Vintraud, M., Zarca, D., Hughes, K. S., Alés Martínez, J. E.: "Clinical practice guidelines for *BRCA1* and *BRCA2* genetic testing"; European Journal of Cancer, 146, 2021, 30–47.

[Schäfer 2023] Schäfer, M. S.: "The Notorious GPT: science communication in the age of artificial intelligence"; Journal of Science Communication, 22(02), 2023, Y02.

[Shafer 1976] Shafer, G.: A Mathematical Theory of Evidence; Princeton: Princeton University Press, 1976.

[Sparks 2011] Sparks, G.: Methods and Tools. Software Development Magazine - Project Management, Programming, Software Testing, 2011.

[Tang et al. 2023] Tang, Y., Zhang, X., Zhou, Y., Huang, Y., Zhou, D.: "A new correlation belief function in Dempster-Shafer evidence theory and its application in classification"; Scientific Reports, 13(1), 2023, 2045–2322.

[Wang et al. 2021] Wang, X.-M., Zhang, X.-R., Li, Z.-H., Zhong, W.-F., Yang, P., Mao, C.: "A brief introduction of meta-analyses in clinical practice and research"; The Journal of Gene Medicine, 23(5), 2021, e3312.

Journal of Universal Computer Science, vol. 29, no. 11 (2023), 1385-1403
submitted: 19/5/2023, accepted: 30/9/2023, appeared: 28/11/2023 CC BY-ND 4.0

Retail Indicators Forecasting and Planning

Nelson Baloian
(Computer Science Department, Universidad de Chile, Santiago, Chile
https://orcid.org/0000-0003-1608-6454, nbaloian@dcc.uchile.cl)

Jonathan Frez
(Computer Science Department, Universidad Diego Portales, Santiago, Chile
https://orcid.org/0009-0003-4812-4483, jonathan.frez@mail.udp.cl)

José A. Pino
(Computer Science Department, Universidad de Chile, Santiago, Chile
https://orcid.org/0000-0002-5835-988X, jpino@dcc.uchile.cl)

Cristóbal Fuenzalida
(Computer Science Department, Universidad Diego Portales, Santiago, Chile
https://orcid.org/0000-0000-0000-0000, c.fuenza6597@hotmail.com)

Sergio Peñafiel
(Fundación Arturo Lopez Perez (FALP), Santiago, Chile
https://orcid.org/0000-0002-0025-7805, sergio.penafiel@falp.org)

Belisario Panay
(Allm Inc., Tokyo, Japan
https://orcid.org/0000-0002-1440-8192, belisariops@gmail.com)

Gustavo Zurita
(Management Control & Information Systems, Universidad de Chile, Santiago, Chile
https://orcid.org/0000-0003-0757-1247, gzurita@fen.uchile.cl)

Horacio Sanson
(Allm Inc., Tokyo, Japan
https://orcid.org/0000-0002-1440-8192, horacio@allm.inc)

Abstract: We present a methodology to handle the problem of planning sales goals. The methodology supports the retail manager to carry out simulations to find the most plausible goals for the future. One of the novel aspects of this methodology is that the analysis is based not on current sales levels, as most previous works do, but on those in the future, making a more precise and accurate analysis of the situation. The work presents the solution for a scenario using three sales performance indicators: foot traffic, conversion rate and ticket mean value for sales, but it explains how it can be generalized to more indicators. The contribution of this work is in the first place a framework, which consists of a methodology for performing sales planning, then, an algorithm, which finds the best prediction model for a particular store, and finally, a tool, which helps sales planners to set realistic sales goals based on the predicted sales. First we present the method to choose the best indicator prediction model for each retail store and then we present a tool which allows the retail manager estimate the improvements on the indicators in order to attain a desired sales goal level; the managers may then perform several simulations for various scenarios in a fast and efficient way. The developed tool implementing this methodology was validated by experts in the subject of administration of retail stores yielding good results.

Keywords: Sales and Operation Planning, Sales Goal Planning, Machine Learning
Categories: H.3.1, H.3.2, H.3.3, H.3.7, H.5.1
DOI: 10.3897/jucs.112556

1 Introduction

A key activity in retail store management is to make a good guess of store sales in the coming days, weeks, months or even years. This is commonly known as Sales and Operation Planning (S&OP) [Thomé, 2012] and is one of the most important tasks managers of retail stores must perform in order to increase store performance. Its results are then used to efficiently plan the provision of resources, such as adequate staffing and enough goods to be sold. According to [Kreuter, 2022] the research in S&OP has been booming in recent years. Sales goal planning is a closely related task and refers to the process of setting targets and objectives for a sales team or individual sales representatives. It involves determining the specific sales objectives, establishing measurable targets, and outlining strategies and actions to achieve those goals. Big data and machine learning based method are increasingly being used to support this task [Aversa, 2021], [Huber, 2020] [Tsoumakas, 2019].

A convenient planning scenario is one in which the manager assumes a target in terms of the expected number of sales. Then, she obtains the values of the store performance indicator allowing her to achieve this target from a certain prediction machine.

There are many typical indicators of store performance (they are reviewed in subsection 2.3). To determine sales and operations planning, managers often use prediction models for these indicators in order to plan their sales goals [Pavlyuchenko, 2021]. The manager evaluates the obtained values of these indicators and may consider them inconvenient or unfeasible. In such a case, she can start asking the machine again with a new sales target as input. The scenario, therefore, implies a simulation cycle, which ends when the administrator is satisfied with the obtained indicator values. Then she can make the decisions necessary to achieve those values. For example, she may set up a promotion and advertising campaign to hopefully achieve the indicator values. If these values actually occur later on and the prediction machine was correct, then the expected sales target would be satisfied.

In this article, we propose a framework, which consists of a methodology for performing sales planning, an algorithm, which finds the best prediction model for a particular store, and a tool, which helps sales planners to perform simulations like the described above, in order to set realistic sales goals based on the predicted sales numbers. In order to validate the framework, we present a real case considering three of the most common performance indicators for retail sales: foot traffic, conversion rate, and average purchase amount per ticket. Therefore, in the mentioned scenario, the output will be a combination of values for these three indicators. These indicators were chosen because they are easy to monitor, yet they convey valuable information.

The reported research is based on using large collections of data and therefore it satisfies one of the topics of this J.UCS special issue.

The rest of the paper is organized as follows: Section 2 presents related work. Section 3 describes our proposal for the Sales Planning Model. Section 4 includes the model implementation and results. Finally, Section 5 presents the conclusions.

2 Related Work

2.1 Sales and Operation Planning

Estimating the outcome of sales promotion has always been a goal for a store chain manager [Boulden, 1957]. Accurate sale estimations allow precise product stocks to be held as well as adequate personnel to be provided. Furthermore, clients being able to purchase the promoted items generate customer satisfaction. On the financial side, accurate sale estimations provide control on cash flow, meet total sales goals and thus obtain expected profits. In short, accurate sales estimation is a core activity for a retail company [Hastings, 1994].

There have been numerous studies on retail indicators forecasting and planning, which involve predicting and preparing for changes in retail sales and supply chain management. These studies have explored various methods including statistical models, judgmental approaches, hybrid methods, data mining techniques, artificial neural networks, support vector machines, random forests, demand forecasting, inventory management, and production planning.

In particular, we are working in an S&OP process, which is a business process used to effectively align organizational demand and supply [Kreuter, 2022]. It involves creating a single plan that integrates sales and marketing, production, and financial goals of an organization.

In S&OP, there is the concept called "demand-driven S&OP", which emphasizes the importance of considering customer demand when developing the S&OP plan [Cassivi, 2007]. Demand-driven S&OP seeks to align the organizational demand and supply by using customer demand as the driving force behind production and inventory plans. This approach is based on the idea that customer demand is the most important factor in determining organizational production and supply chain activities.

The most widely used approaches for demand-driven S&OP correspond to both qualitative and statistical analysis. In qualitative analyses [Garcia-Villareal, 2018], [Danese, 2018], [Hulthén, 2016], experts analyze the behavior of certain key events from which conclusions are obtained about the applied measures. In the statistical and optimization analysis [Nemati, 2013], [Lim, 2014], the goal is to measure the variation of key indicators that are observed when certain measures are applied. This is done through the analysis of historical data or through simulations seeking to replicate the context of the process. It is reported whether the applied measures have a positive impact on the indicators and whether or not this impact is statistically significant to call the set of measures an improvement.

In order to support and improve the demand-driven operation, the use of advanced analytical techniques can be used, such as machine learning and artificial intelligence [Glackin, 2022]. These techniques can be used to analyze historical data and make accurate demand forecasts, which can help organizations better align their demand and supply. In addition, these techniques can be used to identify patterns and trends in demand and supply data, which can help organizations optimize their S&OP processes and make informed decisions.

2.2 Sales Goal Planning

Setting effective sales goals requires a structured approach taking into account the organization sales history, market trends, customer needs, and competitive landscape. Sales goal planning typically includes at least the following steps [Juran, 1992]:

- **Define Objectives:** Start by clearly defining the objectives you want to achieve. These objectives may include revenue targets, market share growth, customer acquisition, or sales volume increase.
- **Analyze Historical Data:** Review past sales performance data to understand trends, identify strengths, weaknesses, opportunities, and threats (SWOT analysis). This analysis supports the setting of realistically achievable goals.
- **Set Specific and Measurable Targets:** Establish specific, measurable, attainable, relevant, and time-bound (SMART) targets. For example, a goal could be to increase sales by 10 percent.
- **Break Down Goals:** Divide the overall sales goal into smaller milestones or benchmarks. This helps create a sense of progress and provides a clear road-map for achieving the final objective.
- **Identify Strategies and Tactics:** Determine the strategies and tactics required to reach the sales goals. This may involve targeting specific customer segments, launching promotional campaigns, improving sales processes, or enhancing product features.
- **Allocate Resources:** Allocate resources such as budget, personnel, training, and technology to support the sales team in achieving their goals. Ensure that the necessary tools and resources are available to maximize their effectiveness.
- **Establish Metrics and KPIs:** Define key performance indicators (KPIs) and metrics that will be used to track progress and measure success. Common sales metrics include revenue, conversion rates, average deal size, customer acquisition cost, and customer retention rate.
- **Monitor and Evaluate Progress:** Continuously monitor sales performance against the set goals and KPIs. Regularly review progress and identify any areas where adjustments or improvements are needed.

Sales goal planning is an iterative process that requires ongoing monitoring, evaluation, and adaptation to ensure continued success.

Several sales goal planning frameworks have been proposed to help organizations set and achieve their sales targets. One popular framework is the Objective and Key Results (OKR) framework [Al Thinyan, 2022]. The OKR framework involves setting ambitious but achievable objectives and defining measurable key results to track progress towards those objectives. Another popular framework is the Sales Pipeline Management Framework [Park, 2020], which involves identifying and tracking the various stages of the sales process, from lead generation to closing the deal. The Sales Pipeline Management Framework helps organizations to identify bottlenecks in the sales process and optimize their sales strategies to improve conversion rates and sales speed.

2.3 Retail Sales Performance Indicators

When it comes to measuring and evaluating the performance of a retail sales operation, there are several key performance indicators (KPIs) that can provide valuable insights [Anand, 2015]. Some of them are the following:

- **Sales Revenue:** The total amount of revenue generated from retail sales within a specific period, such as a day, week, month, or quarter.
- **Sales Growth:** The percentage increase or decrease in sales revenue compared to a previous period. It indicates the overall growth rate of retail sales.
- **Average Transaction Value (ATV):** The average amount of money spent by customers in a single transaction. Calculated by dividing the total sales revenue by the number of transactions.
- **Conversion Rate:** The percentage of potential customers who make a purchase. It is calculated by dividing the number of transactions by the number of visitors or footfall in a store.
- **Customer Traffic:** Also known as foot traffic, it refers to the total number of visitors or footfall in a retail store during a given period. It helps assess the store popularity and customer attraction.
- **Customer Acquisition Cost (CAC):** The average cost associated with acquiring a new customer. It includes marketing and advertising expenses, sales commissions, and other related costs.
- **Gross Margin:** The difference between the total sales revenue and the cost of goods sold (COGS). It represents the profitability of the products being sold.
- **Inventory Turnover:** The number of times inventory is sold or replaced during a specific period. It indicates how efficiently a retailer is managing inventory and can help identify slow-moving or obsolete items.
- **Return on Investment (ROI):** The ratio of the net profit generated from retail sales to the total investment made. It helps evaluate the profitability and efficiency of the retail operation.
- **Customer Satisfaction:** Measuring customer satisfaction through surveys, feedback, or ratings can provide insights into the overall customer experience and loyalty. Satisfied customers are more likely to repeat purchases and recommend the store to others.

These are just a few examples of the many performance indicators that can be utilized in retail sales. The specific chosen KPIs will depend on the nature of the retail business, its goals, and the available data. It is important to regularly track and analyze these indicators to identify areas of improvement and make informed decisions to optimize sales performance.

2.4 Sales Goal Planning and Machine Learning

In the literature, we can find success stories in which the use of machine learning has helped improve retail indicators. For example, one of the simplest but widely used methods is ARIMA, which is a model based on moving averages which is capable of interpolating and extrapolating the behavior of a time series. Lam et al. [Lam, 2020] use ARIMA to forecast store foot-traffic; with this information, they propose a strategy to find the optimal retail sales force.

McIntyre et al. [McIntyre, 1993] propose an early use of Artificial Intelligence to the problem of estimating the outcome of promotional sales. They mention ten factors explaining why making estimation based just on human experience is a very difficult and error-prone process. Therefore, they propose case-based reasoning as a computer-supported tool intended to incorporate the human expertise within the firm. It is an example of using an Organizational Memory [Guerrero, 2001]. They choose ten factors to be included in the system, such as season of the promotion, number of television spots and percentage of price discounts. Guo et al. [Guo, 2013] developed a multivariate intelligent decision making model. The corresponding system is composed of three modules: a data preparation and preprocessing module, a harmony search-wrapper-based variable selection module, and a multivariate intelligent forecaster module. The forecasting precision of the model was tested as better than other proposals at that time (2013). Machine learning appears as an interesting approach to forecast sales. For example, the manager does not need to select the factors to be included in the prediction system, as the case-based reasoning tool mentioned above [McIntyre, 1993] required. This occurs because the machine learning system chooses the best factors. Bendle et al. [Bendle, 2021] have reviewed the issues and perspectives of machine learning in the retailing area.

Another example is the one presented by Panay et al. [Panay, 2021] in which a model is presented that can predict the behavior of foot traffic, conversion rate and sales amount in a range of time in the future. The method uses the historical data of these indicators to train a custom regression model. The model achieves high accuracy in all variables, with an average RMSE of 0.07. In addition, the model is capable of indicating the most relevant variables at the time of making the forecast. However, although this model allows obtaining precise results on the behavior of sales, it does not provide optimal strategy planning to maximize sales.

Conflicting variables involved in promotional planning imply visualizing these effects. Stewart and Gallen [Stewart, 1998] proposed a matrix to visualize such conflicts in budget allocations.

2.5 Proposed Approach and its Relation to Previous Work

We propose a framework consisting of a model and a concrete tool based on it to support the first steps of a sales goal planning process. Therefore, we use the general context described in subsection 2.1. The basic steps for the process are assumed to be the ones listed in subsection 2.2. The KPIs (subsection 2.3) we choose are foot traffic, conversion rate, and average transaction value, which are well-known indicators. For the prediction of the indicators, we use the model by Panay et al. [Panay, 2021] (subsection 2.4). Nevertheless, we extend the applicability of this model, by providing automatic strategy planning to maximize sales. Finally, planning results must be presented in a suitable way; we inspired in Stewart and Gallen [Stewart, 1998] work (subsection 2.4) to offer visual presentation of the results.

3 The Sales Goal Planning Model

Our proposal for supporting the sales goal-planning process is a tool, which sales managers can use to simulate various scenarios. With this tool, they can receive suggestions about the necessary changes that they have to achieve in the values of the three sales performance indicators chosen for this work (foot traffic, conversion rate, and average transaction value) in order to attain a particular total sales amount. In this way, they can explore various scenarios for deciding which are the most convenient actions to take for achieving the suggested values. For this purpose, managers can organize increasing advertising and/or sales promotion campaigns. Unlike most procedures reported in the literature for sales goal planning, this proposal does not use current values for the three KPI's as starting values; instead, it uses values obtained by a prediction model for the period being simulated. This strategy avoids the effects of stationary variations and/or increasing or decreasing trends. The next subsection describes the prediction model and the following one, the way how they are used in order to explore recommendations for the managers to modify the values of foot traffic, conversion rate and average transaction value to achieve the desired sales goal.

3.1 Prediction and Model Selection Process

In a previous work [Panay, 2021], we realized that a predictive model could perform well for certain stores in particular, which was complemented with hyper-parameterization of the model and search for the most appropriate hyperparameters for each store. However, through experimentation, we discovered that some models perform better than others with the correct hyperparameters. Therefore, in this work we designed an algorithm that explores the use of various models with a constrained set of hyperparameters, choosing the one that provides the best results and then improving the selection of the hyperparameters of the chosen model.

The selection process tests the following estimator algorithms:

- **Gradient Boosting Decision Trees (GBDT)** is an algorithm that combines multiple decision tree models to create a more powerful model for prediction. It works by training a decision tree model on the data and then iteratively improving the model by adding new decision trees that focus on the mistakes made by the previous trees. Each new tree is trained using the residual errors from the previous tree, which helps to correct the mistakes and improve the overall accuracy of the model.

Support Vector Regression (SVR) is an algorithm used to predict continuous numerical values. It works by finding the optimal hyperplane in a higher-dimensional space that maximally separates the data into two classes. The hyperplane is chosen such that the distance between it and the closest data points is maximized, which is known as the margin. SVR then uses this hyperplane to make predictions for new data points. SVR can handle both linear and nonlinear relationships between the predictors and the target variable.

XGBoost Regression (XGBR) is an algorithm that is used to predict continuous numerical values. It is a type of Gradient Boosting algorithm which is specifically designed for regression problems. XGBR works by training a series of decision tree models and using the predictions of these trees to improve the overall accuracy of the model. It is known for its ability to handle large datasets, high-dimensional data, and missing values.

Random Forest is an algorithm that is used to predict both continuous numerical values (regression) and discrete categories (classification). It works by training multiple decision tree models on random subsets of the data and then aggregating the predictions of these trees to make a final prediction. Each tree is trained using a different subset of the data and a different subset of the features, which helps to reduce overfitting and improve the overall accuracy of the model. Random Forest can handle both linear and nonlinear relationships between the predictors and the target variable.

Weighted Evidence Regression Model (WEVREG) is an interpretable regression method where the prediction of a new observation can be easily followed [Panay, 2020]. This model is based on a previous one that was proposed by Petit-Renaud and Denœux [Petit-Renaud, 2004] It uses the Dempster–Shafer Theory together with a variation of K-Nearest Neighbors (KNN) to produce an evidence-based regression model. WEVREG extends this model by adding weights to each dimension of the attributes, making certain parameters have more importance than others when deciding the most similar instances of the KNN process. In addition, this change allows measuring the importance of each attribute in the prediction.

Calibrating the hyperparameters of an estimator algorithm is necessary to achieve the best possible fit to the data. However, because each store may have unique characteristics, it is not feasible to use the same estimator with the same hyperparameters for all stores. Instead, we must calibrate the hyperparameters for each store separately to ensure optimal performance.

Finding the optimal hyperparameters for an estimator algorithm and a store can be a time-consuming task, as it involves testing all possible combinations of these parameters. To reduce the amount of time required, we can limit the search by defining ranges and the number of values to test for each parameter. For example, if we want to find the best value for a parameter i, we might define a range of [0-1000] and test 10 values within that range (0, 100, 200, …, 1000). However, most estimators have more than one parameter, so we might also need to define ranges and values for additional parameters, such as j. This process of testing all possible combinations of multiple parameters is known as Exhaustive GridSearch.

We can follow these steps to make the GridSearch process efficient:

- Conduct GridSearch for each estimator and store, using the same ranges and values for all stores. Save the selected points from each grid.

- Count the number of times each point is selected across all stores and estimators. Select the K most common points, creating a new grid with these K combinations of hyperparameters. This process will result in a set of 5 grids, each containing the K most common hyperparameter combinations that performed best during the estimator training process.

The goal at this stage is to choose the estimator that will give us the lowest margin of error. We can achieve this goal by training and testing a model for each estimator using the most common hyperparameter from the corresponding MaxGrid. Alternatively, we can also use the N most common hyperparameters from each MaxGrid to increase our chances of selecting the best one.

After we have the margin of error for each estimator, we select the one that performs the best and evaluate the full MaxGrid with it. Afterwards, we can choose the optimal set of hyperparameters and use it to train our final model with the selected estimator.

The outcome of this training and algorithm selection process enables us to reach a significantly higher number of stores on the platform, with much more accurate predictions than those generated manually by analysts and resource planners. This results in the ability to utilize these insights for the development of effective commercial strategies and resource planning. The main steps of the algorithm are described in Algorithm 1.

Algorithm 1 Hyperparameter Calibration for Store-wise Estimator Algorithm (details)

Input: estimators, stores, ranges, values, K, N **Output:** Optimal hyperparameters and estimator

```
Initialize empty list $MaxGrids$
for each estimator $e$ and store $s$ do
    Conduct GridSearch with ranges and values
    Save selected points from grid in list $selected$
    for each point $p$ in $selected$ do
        Increment count of $p$ in a dictionary $counts$
    end for
end for
for each dictionary $counts$ do
    Select the K most common points, add to a new grid
    Append the new grid to the list $MaxGrids$
end for
for each estimator $e$ and grid $g$ in $MaxGrids$ do
    Train and test a model with the most common hyperparameter from $g$
    Record the margin of error
end for
Choose the estimator with the lowest margin of error
Evaluate the full MaxGrid with the chosen estimator
Choose the optimal set of hyperparameters and train the final model with the chosen estimator
```

3.2 Sales Goal Setting Model

Once the model is working and giving results with a desirable precision, the next step is to bring it together with available historical data from stores. This procedure will allow us to build a "sales-goal planning tool" on the user-end, which allows managers to see both the predictions for their store indicators, and the increment of those values needed to achieve their sales goals.

To carry out the aforementioned purpose, we created a model that takes the predicted indicators as an input and gives the associated variations needed as outputs in function of the coefficients of variation of indicators obtained from the historical data; the latter being representative of how easily the indicators can vary. To simplify their use, we work with their reciprocals, which we will call weights for the rest of the explanation, denoted as WX. The model is defined as follows:

- We define the amount of sales *S* as the product between the number of enters (foot traffic) *E*, the conversion rate *C* and the average purchase amount per ticket *T*.

$$S = E \times C \times T \tag{1}$$

- The predicted indicators (inputs) will have the subscript *p*, and the indicator goals (outputs) will have the subscript *g* (note that S_g is the sales goal selected by the user and is therefore a constant). Thus we have:

$$S_p = E_p \times C_p \times T_p \tag{2}$$

$$S_g = E_g \times C_g \times T_g \tag{3}$$

- Since the percentage change of these indicators are the focus of the algorithm, we use the following definitions to proceed:

$$V_S = \frac{S_g}{S_p} - 1 \quad V_E = \frac{E_g}{E_p} - 1 \quad V_C = \frac{C_g}{C_p} - 1 \quad V_T = \frac{T_g}{T_p} - 1 \tag{4}$$

- Replacing those definitions in (3), we get:

$$S_p(1 + V_S) = E_p(1 + V_E) \times C_p(1 + V_C) \times T_p(1 + V_T)$$

$$\Rightarrow [E_p \times C_p \times T_p](1 + V_S) = [E_p \times C_p \times T_p](1 + V_E)(1 + V_C)(1 + V_T)$$

$$\Rightarrow (1 + V_S) = (1 + V_E)(1 + V_C)(1 + V_T) \tag{5}$$

- As mentioned before, the weights of indicators represent the reciprocal quantities of how easily they can vary, which means variations and weights must be inversely proportional to each other:

$$V_E \cdot W_E = V_C \cdot W_C = V_T \cdot W_T = x \qquad \text{(proportion ratio)} \tag{6}$$

- Then, without loss of generality, we will start to approach the problem focusing on one of the variables, because of the symmetry of the product that defines the equation. We will use the equation (5) and will leave it as a function of V_E; taking V_S and the weights as constants (because their values are fixed for the problem).

$$\Rightarrow (1 + V_s) = (1 + V_E)\left(1 + V_E * \frac{W_E}{W_C}\right)\left(1 + V_E * \frac{W_E}{W_T}\right) \tag{7}$$

- When rearranging the equation and amplifying it by $(W_E \cdot W_C \cdot W_T)$, we get the following third degree polynomial equation:

$$(V_E \cdot W_E)^3 + (W_E + W_C + W_T)(V_{E\cdot} \cdot W_E)^2 +$$
$$\left(\frac{1}{W_E} + \frac{1}{W_C} + \frac{1}{W_T}\right)(W_E \cdot W_C \cdot W_T)(W_E \cdot W_E) + \tag{8}$$
$$(W_E \cdot W_C \cdot W_T) \cdot (-W_S) = 0$$

- To synthesize the notation, we define the following constants in function of the weights:

$$c_1 = W_E + W_C + W_T$$

$$c_2 = \frac{1}{W_E} + \frac{1}{W_C} + \frac{1}{W_T}$$

$$c_3 = W_E \cdot W_C \cdot W_T$$

$$c_4 = -V_S$$

- Then, we replace those values in equation (8) and use the proportion $x = V_E \cdot W_E$ from (6), obtaining the following cubic polynomial equation:

$$x^3 + c_1 \cdot x^2 + c_2\, c_3 \cdot x + c_3\, c_4 = 0 \tag{9}$$

- The previous equation no longer depends specifically on the indicator and weight associated with foot traffic (*E*), but now depends entirely on the proportion met by the three indicators and the constants c_k. It can be easily seen that the problem is generalized for the three indicators in the same way, since the original equation is symmetric. Therefore, the whole problem boils down to solving the cubic equation as a function of *x*, for which the general formula of Gerolamo Cardano is used.

Finally, having the proportion ratio x, we can figure the variations using (6) and from there the indicator goals needed for the desired sales goal using the definitions from (4), concluding the algorithm.

Note that similarly, if the user had chosen to fix one or two of the indicators as constants, the problem would turn into second and first order polynomial equations respectively, being easier to solve in the same way.

3.3 Multi Variable Optimization

The mathematical model provides a framework to determine the necessary adjustments required to reach a sales target given these three variables in our system. However, there are other variables outside the system that can be used in combination with the predicted Foot traffic, Conversion rate and Sales. For example, Staffing, Inventory and Customer Satisfaction rates. Assuming that equation (1) can be expanded to include more variables, we can incorporate them into the optimization process to achieve the sales goal.

For example, in (10), V is the average value of other relevant variable(s)

$$S = E \times C \times T \times V \tag{10}$$

In this particular scenario (10), we introduce a single variable, which is directly proportional to sales. An example of such a variable could be staffing. However, in order to allow for the flexibility of modifying the sales modeling function with more complex options that incorporate additional variables and alternative forms of relationships, an objective function will be employed. This objective function aims to determine the optimal values, regardless of the specific form of the sales function being used. The objective function measures the deviation between the actual sales (S) and the sales goal (G). This function can be defined as the squared difference between S and G (11), and with this function the optimization process can find the optimal values for the variables E, C, T, and V in order to achieve the desired sales goal.

$$\text{diff} = (\text{S - G})^2 \tag{11}$$

We employed Bayesian optimization to obtain the optimized values for each variable. This optimizer systematically evaluates the objective function by sampling various points in the search space. Its aim is to minimize the squared difference between the actual sales and the sales goal. By exploring various combinations of E, C, T, and V values, it searches for the optimal solution.

For instance, consider having a foot traffic of 218, an average ticket value of \$140,238, a conversion rate of 20.18%, resulting in \$6,170,189 of sales. Our sales goal is set at \$7,500,000. Additionally, let us introduce another variable called staffing, which has a value of 1.

Through Bayesian optimization, we can find the best combination of E, C, T, V, and staffing values that minimize the difference between the achieved sales and the desired sales goal. This approach allows us effectively explore the search space and identify the optimal configuration for maximizing sales performance.

Based on the optimization results in this example, it is suggested that the foot traffic should be increased to 221, the conversion rate should be improved to 21%, and the average ticket value should be set to $145,854. However, it is important to note that there is no need to hire additional staff based on the optimization outcome. By making these adjustments to the variables, it is expected that the sales goal can be achieved without the need for additional staff. This information provides valuable insights on how to optimize the sales process and drive revenue growth effectively.

Also, it is important to note that the equation (10) and the optimization function (11) in combination with a Bayesian optimization, allows us add as many variables as we need to the model in a simple and easy to understand way. A general overview of the method is described in Algorithm 2.

Algorithm 2 running the Bayesian optimization over the sales function

Input: Ep, Cp, Tp, and V_avg[]**, Output:** optimized values of S, E, C, T, and V[]

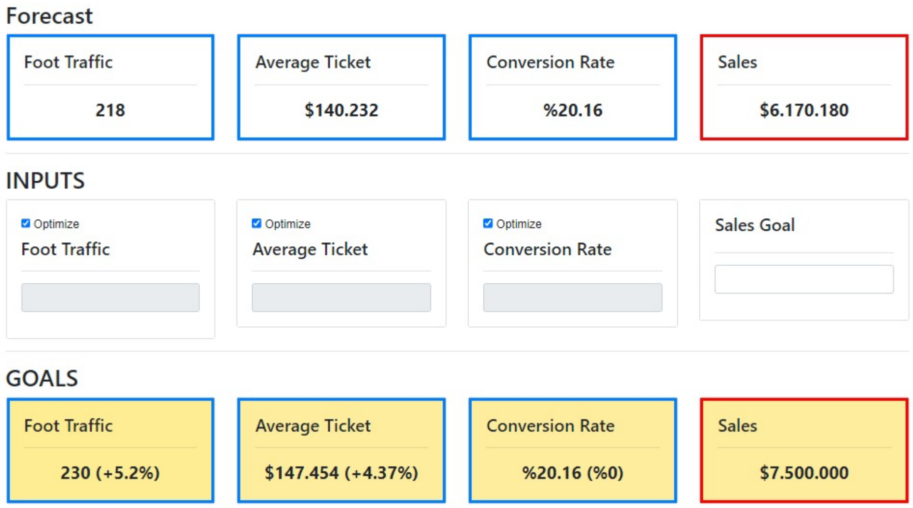
```
1: define the **sales_function**(Ep, Cp, Tp, V_avg[]) as a function returning an estimation of sales based on the predicted values of E, C, T and a varying set of average values.
2: define the **sales_goal_objective** that calculates the squared difference between the sales amount and the sales goal using the **sales_function**
3: Perform the Bayesian optimization using the **sales_goal_objective** function
4: **Calculate** the optimized sales amount So using the optimized values Eo, Co, To, Vo_avg[])
5: return **So, Eo, Co, To, Vo_avg[]**
```

4 Implementation and Results

The implementation of the sales-goal-planning tool has been met with success in both user acceptance and performance since its introduction for companies' monthly goal planning. It currently boasts a prediction accuracy of over 90%, however, we need to compare these results with the previous human performance accuracy.

To facilitate the user-end implementation of the planning tool, we developed a web-app which collects and displays all results. The app includes various fields for user input, such as desired sales goals and optional indicators. Users can also select the desired date range for predictions, with the default being two months in advance. The resulting goals and percentage change needed to achieve them are displayed prominently on the app.

As seen in the accompanying images, the flexibility of the tool is demonstrated through two examples. In the first, the user only inputs their desired sales goal (see Figure 1). In the second, they also set the foot traffic value as a constant for the algorithm (see Figure 2). It is worth noting that fixing the foot traffic near the predicted value results in larger variations for other indicators, making them harder to achieve.



Forecast

Foot Traffic
218

Average Ticket
$140.232

Conversion Rate
%20.16

Sales
$6.170.180

INPUTS

☑ Optimize
Foot Traffic

☑ Optimize
Average Ticket

☑ Optimize
Conversion Rate

Sales Goal

GOALS

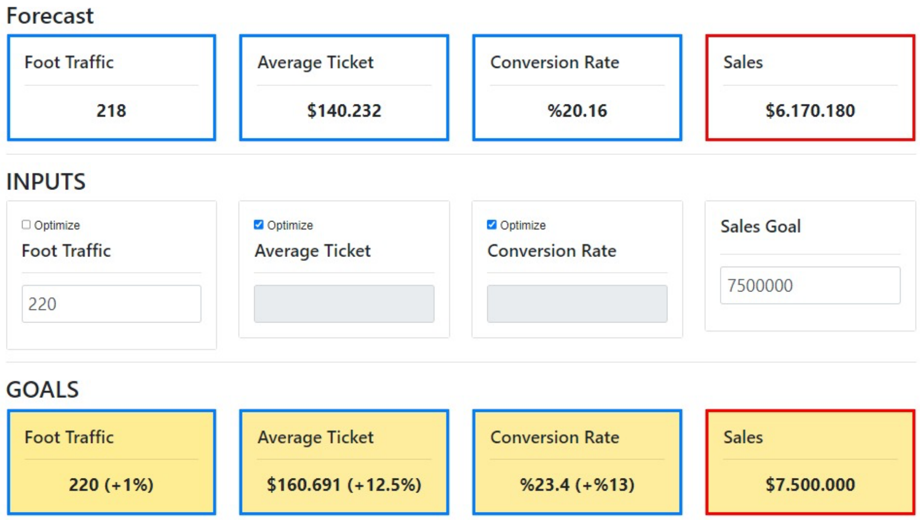
Foot Traffic
230 (+5.2%)

Average Ticket
$147.454 (+4.37%)

Conversion Rate
%20.16 (%0)

Sales
$7.500.000

Figure 1: Sales-goal-planning tool results with only Sales Goal as input



Forecast

Foot Traffic
218

Average Ticket
$140.232

Conversion Rate
%20.16

Sales
$6.170.180

INPUTS

☐ Optimize
Foot Traffic
220

☑ Optimize
Average Ticket

☑ Optimize
Conversion Rate

Sales Goal
7500000

GOALS

Foot Traffic
220 (+1%)

Average Ticket
$160.691 (+12.5%)

Conversion Rate
%23.4 (+%13)

Sales
$7.500.000

Figure 2: Sales-goal-planning tool results with Sales Goal as input

In order to assess the effectiveness of the tool, we compared its performance against the previously used estimations. These estimations were conducted by a team of experts employed by the holding company, tasked with analyzing the performance of each store across various countries. This intricate process involves collating information from store managers and integrating it with data available in the system. As a result, the outcome represents a collaborative endeavor of the analytics unit. This sum of efforts is referred to as "Human Performance" (HP). A set of 30 stores estimations were selected for this comparison. A team of two to three experts worked on the estimation for each store. They were the people responsible for this task up to now. This allowed us to gauge the accuracy of the tool and its efficiency compared to human performance. The results of a comparison between Human Performance labelled as "HP" and the tool labeled as "Algorithm", are shown in Table 1. The "average" row represents the average error value achieved by each method. We can see that the HP method achieved an average value of -7.73% and the Algorithm method achieved an average error value of -4.44%.

	HP	**Algorithm**		**HP**	**Algorithm**
store1	-14%	-23%	store16	-17%	-8%
store2	-12%	-20%	store17	-21%	-8%
store3	-19%	-20%	store18	-9%	-8%
store4	-26%	-19%	store19	-18%	-7%
store5	-13%	-17%	store20	5%	-5%
store6	-22%	-16%	store21	-20%	-4%
store7	-7%	-16%	store22	-1%	-3%
store8	-16%	-14%	store23	-3%	-2%
store9	-25%	-13%	store24	-9%	-1%
store10	-12%	-13%	store25	-9%	-1%
store11	-15%	-12%	store26	4%	2%
store12	-3%	-11%	store27	4%	2%
store13	-28%	-11%	store28	-12%	2%
store14	-11%	-10%	store29	0%	2%
store15	-8%	-8%	store30	5%	6%
			HP	**Algorithm**	
		average	-7.73%	-4.44%	
		st. deviation	0.516	0,507	
		min error	-1%	-1%	
		max error	-28%	-23%	

Table 1: Monthly sales estimation error for each store and statistics. Negative and Positive values refer to sub estimations or over estimations of foot traffic

The “min error” row represents the lowest error value achieved by each method, which was the same: -1%. The “max error” row represents the highest error value achieved by each method; the HP method achieved a maximum value of -28% and the Algorithm method achieved a maximum value of -23%.

The “st. deviation” row represents the standard deviation of the results for each method, which is a measure of the variability or dispersion of the data. It can be seen that the standard deviation for the HP method is 0.516 and for Algorithm method is 0.507.

In certain instances, it is relevant to highlight that the disparity between HP and the Algorithm is more pronounced than in others. This discrepancy can be attributed to the fact that predictions are not uniformly derived from the same dataset. The experts’ additional insights about the environment, such as improvements in window displays or enhanced local management of the establishment, are not necessarily reflected in the data. For example, in the cases of stores 1, 2, 7 and 12 the experts’ estimation (HP) yields clearly superior outcomes. Conversely, for stores 9,16, 17, 19, 21, 24, 25 and 28 the algorithm demonstrated a heightened level of precision. We attribute this latter contrast to an underestimation of the chain’s marketing campaign, which can be observed through the consistent number of negative predictions across nearly all stores.

The results show the Algorithm method has similar results in terms of min, max, avg and st. deviation (slightly better for the Algorithm).

Moreover, as the reader may have already noticed, the times the algorithm outperforms HP notably are more prevalent—4 instances compared to 7. This underscores the algorithm's value as an option that excels at conducting a comprehensive analysis across a large number of stores using shared metrics when a broad assessment is required. However, as we pointed out before, the experts may have deep knowledge about few particular stores and therefore, their predictions for these stores may be very precise.

As the model was being developed, an implementation of it was created in the form of a web-app to enhance the approach used by multiple retailers and business intelligence companies in providing feedback to their clients, many of whom are among the largest retail companies in Chile and Colombia. After utilizing this implementation for over a year, we sought to gather their perspectives on its usefulness, the reasons for seeking out a new solution, and their future projections for its use.

The positive news is that they continue to utilize the tool, consistently incorporating dozens of additional stores into the system every month. They have found it to be a valuable asset in enhancing their ability to accurately predict sales and set goals for each store.

5 Conclusions

This work presents a methodology to support retail managers while performing sales and operations planning activities. The system consists of two parts: an algorithm for forecasting store performance indicators (foot traffic, conversion rate, and average transaction value) using evidential regression, and a method for planning sales goals based on these predictions. The algorithm is designed to provide a clear interpretation of the prediction process by identifying the most important features for the forecast. The method for planning sales goals uses the predictions and historical data on the variation of the indicators to find the optimal daily variations that are most likely to achieve the sales goal and associated risk according to the users (planners). Also, the quantitative analysis indicates that the Algorithm method has slightly better prediction performance compared to the experts' estimations. However, the results of Algorithm method are similarly consistent with the experts' results.

The described model and tool have been implemented in the real world. They are being actually used in stores chains in Chile and Colombia for over a year with positive results.

The reported research should be of interest to managers in charge of the Sales and Operation Planning of brick and mortar stores. As we mentioned, this planning is key for the performance of those stores. It may be even more relevant at present time in which the brick and mortar retail business has a strong challenge from online retail commerce.

Future research can be done in several areas. One of them is to try to make the Algorithm method results more consistent than they are now. Another area for research is to explore other performance indicators which may make easier and/or performing better the manager's task than it is now.

References

[Al Thinyan, 2022] Al Thinyan, K., Ghawji, H., Al Shehri, A.: What are OKRs and KPIs and can they Coexist within an Organization. International Journal of Innovative Science and Research Technology 7(8), 176-179, 2022.

[Anand, 2015] Anand, N., Grover, N.: Measuring retail supply chain performance: Theoretical model using key performance indicators (KIPs). Benchmarking: An International Journal 22(2), 290-308, 2015.

[Aversa, 2021] J. Aversa, J., Hernandez, T., Doherty, S.: Incorporating big data within retail organizations: A case study approach. Journal of Retailing and Consumer Services 60,102447, 2021.

[Bendle, 2021] Bendle, N., Wang, P.X., Ryoo, J.: The role of machine learning analytics and metrics in retailing research. Journal of Retailing 97, 658–675, 2021.

[Boulden, 1957] Boulden, J.B.: Fitting the sales forecast to your firm. Business Horizons 1, 65–72, 1957.

[Cassivi, 2007] Cassivi, H.P.: The role of joint collaboration planning actions in a demand-driven supply chain. Industrial Management & Data Systems 107(7), 954-78, 2007.

[Danese, 2018] Danese, P., Molinaro, M., Romano, P.: Managing evolutionary paths in sales and operations planning: key dimensions and sequences of implementation. International Journal of Production Research 56(5), 2036–2053, 2018.

[Garcia-Villareal, 2018] Garcia-Villarreal, E., Bhamra, R., Schoenheit, M.: A framework for technology selection to support sales and operations planning in German medical technology organisations. Advances in Transdisciplinary Engineering 285, 2018.

[Glackin, 2022] Glackin, C.E., Adivar, M.: Using the power of machine learning in sales research: process and potential. Journal of Personal Selling & Sales Management 26,1-7, 2022.

[Guerrero, 2001] L.A. Guerrero and J.A. Pino. Understanding Organizational Memory. SCCC 2001: 21st International Conference of the Chilean Computer Science Society, 124-132, IEEE, 2001.

[Guo, 2013] Guo, Z.X., Wong, W.K., Li, M: A multivariate intelligent decision making model for retail sales forecasting. Decision Support Systems 55, 247–255, 2013.

[Hastings, 1994] Hastings R. Fildes, R.: The organization and improvement of marketing forecasting. The Journal of Operational Research Society 45, 1–16, 1994.

[Huber, 2020] Huber J., Stuckenschmidt, H.: Daily retail demand forecasting using machine learning with emphasis on calendric special days. International Journal of Forecasting 36(4), 1420–1438, 2020.

[Hulthén, 2016] Hulthén, H., Näslund, D., Norrman, A.: Framework for measuring performance of the sales and operations planning process. International Journal of Physical Distribution & Logistics Management 46(9), 809-835. 2016.

[Juran, 1992] Juran, J.M.: Juran on quality by design: the new steps for planning quality into goods and services. Simon and Schuster, 1992.

[Kreuter, 2022] Kreuter, T., Scavarda, L.F., Thomé, A.M.T., Hellingrath, B., Seeling, M.X.: Empirical and theoretical perspectives in sales and operations planning. Review of Managerial Science 16(2), 319–354, 2022.

[Lam, 1998] Lam, S., Vandenbosch, M., Pearce, M.: Retail sales force scheduling based on store traffic forecasting. Journal of Retailing 74(1), 61–88, 1998.

[Lim, 2014] Lim, L., Alpan, G., Penz, B.: Reconciling sales and operations management with distant suppliers in the automotive industry: A simulation approach. International Journal of Production Economics 151, 20–36, 2014.

[McIntyre, 1993] McIntyre, S., Achabal, D., Miller, C.: Applying case-based reasoning to forecasting retail sales. Journal of Retailing 69(4), 372–398, 1993.

[Nemati, 2017] Nemati, Y., Madhoshi, M., Ghadikolaei, A.S.: The effect of sales and operations planning (S&OP) on supply chain's total performance: A case study in an Iranian dairy company. Computers & Chemical_Engineering 104, 323–338, 2017.

[Panay, 2020] Panay, B., Baloian, N., Pino, J.A., Peñafiel, S., Sanson, H., Bersano, N.: Feature selection for health care costs prediction using weighted evidential regression. Sensors 20(16), 4392, 2020.

[Panay, 2021] Panay, B., Baloian, N., Pino, J.A., Peñafiel, S., Frez, J., Fuenzalida, C., Sanson, H., Zurita, G.: Forecasting key retail performance indicators using interpretable regression. Sensors 21(5), 1874, 2021.

[Pavlyuchenko, 2021] K. Pavlyuchenko, K., Panfilov, P.: Application of predictive analytics to sales planning business process of fmcg company. In Procs. of the 13th International Conference on Management of Digital EcoSystems, 167–170, 2021.

[Park, 2020] Park, K., Lee, G., Kim, C., Kim, J., Rhie, K., Lee, W.B.: Comprehensive framework for underground pipeline management with reliability and cost factors using Monte Carlo simulation. Journal of Loss Prevention in the Process Industries 63,104035, 2020.

[Petit-Renaud, 2004] S. Petit-Renaud S., Denœux, T.: Nonparametric regression analysis of uncertain and imprecise data using belief functions. International Journal of Approximate Reasoning, 35(1), 1–28, 2004.

[Stewart, 1998] Stewart D., Gallen, B.: The promotional planning process and its impact on consumer franchise building: the case of fastmoving goods companies in New Zealand. Journal of Product Brand Management 7, 557–567, 1998.

[Thomé, 2012] Thomé, A.M.T., Scavarda, L.F., Fernandez, N.S., Scavarda, A.J.: Sales and operations planning: A research synthesis. International Journal of Production Economics 138(1), 1–13, 2012.

[Tsoumakas, 2019] Tsoumakas, G.: A survey of machine learning techniques for food sales prediction. Artificial Intelligence Review 52(1), 441–447, 2019.

Journal of Universal Computer Science, vol. 29, no. 11 (2023), 1404-1421
submitted: 19/5/2023, accepted: 30/9/2023, appeared: 28/11/2023 CC BY-ND 4.0

Efficiently Finding Cyclical Patterns on Twitter Considering the Inherent Spatio-temporal Attributes of Data

Claudio Gutiérrez-Soto
(Departamento de Sistemas de Información,
Universidad del Bío-Bío, Concepción, Chile
https://orcid.org/0000-0002-7704-6141, cogutier@ubiobio.cl)

Patricio Galdames
(Facultad de Ingeniería, Arquitectura y Diseño
Universidad San Sebastián, Concepción, Chile
https://orcid.org/0000-0003-3051-2413, patricio.galdames@uss.cl)

Daniel Navea
(Universidad del Bío-Bío, Concepción, Chile
daniel.navea1601@alumnos.ubiobio.cl)

Abstract: Social networks such as Twitter provide thousands of terabytes per day, which can be exploited to find relevant information. This relevant information is used to promote marketing strategies, analyze current political issues, and track market trends, to name a few examples. One instance of relevant information is finding cyclic behavior patterns (i.e., patterns that frequently repeat themselves over time) in the population. Because trending topics on Twitter change rapidly, efficient algorithms are required, especially when considering location and time (i.e., the specific location and time) during broadcasts. This article presents an efficient algorithm based on association rules to find cyclical patterns on Twitter, considering the inherent spatio-temporal attributes of data. Using a Hash Table enhances the efficiency of this algorithm, called HashCycle. Notably, HashCycle does not use minimum support and can detect patterns in a single run over a sequence. The processing times of HashCycle were compared to the Apriori (which is a well-known and widely used on diverse platforms) and Projection-based Partial Periodic Patterns (PPA) algorithms (which is one of the most efficient algorithms in terms of processing times). Empirical results from two spatio-temporal databases (a synthetic data set and one based on Twitter) show that HashCycle has more efficient processing times than two state-of-the-art algorithms: Apriori and PPA.

Keywords: Cyclical Patterns, Spatio-temporal Data, Social Networks, Algorithms
Categories: H.3.1, H.3.2, H.3.3, I.1.2, I.5.3
DOI: 10.3897/jucs.112523

1 Introduction

Data science is a powerful tool to discover hidden knowledge from several sources such as mobile applications [Rejeb et al., 2022], Internet of Things (IoT) [Molinaro and Orzes, 2022], Industry 4.0 [Bhattacharya et al., 2022], and Smart Cities [Kandt and Batty, 2021]. These applications involve sources such as Global Positioning Systems (GPS) [Li et al., 2019] and Geographic Information Systems (GIS) [Lü et al., 2019], which are usually managed by Spatio-Temporal Databases (STDB). STDB stores and processes Spatio-temporal data. These databases are well-known because they incorporate the temporal dimension as an additional attribute, which provides an extra complexity to be processed. Because of its nature, STDB tends to be extremely large, and simple transactions such as insertions and updates become expensive in terms of processing time [Gaede and Günther, 1998]. As a result, traditional Data Mining techniques must be adapted to incorporate the temporal attribute. Existing studies have demonstrated that the analysis of Spatio-temporal databases can yield valuable patterns such as patterns of animal movement, disease response, weather analysis [Revesz, 2009], climate change [Li et al., 2011], road traffic [Nakata and Takeuchi, 2004], criminal activity or terrorist events [Bora et al., 2013], and human sociological behaviour in location-based social networks [Song et al., 2015]

Many patterns can be extracted from STDB, particularly sequential patterns where events or episodes occur over time in a specific sequence. Broadly speaking, there are many types of patterns; however, among the best-known patterns, we can point out periodic patterns. Periodic patterns correspond to those that appear with certain regularity within a sequence. These patterns occur in specific time periods. For instance, a periodic pattern corresponds to the schedule in which the bus "Line 1" arrives at Grand Central Station in Manhattan. This schedule can be given each hour from 6:00 to 18:00. From this baseline, a period can represent any unit of time, therefore a period can be an hour, a day, a week, and so on. On the other hand, cyclical patterns are an extension of periodic patterns, where the events that make up a pattern can be given by an interval of periods. For example, in the peak hours between 18:00 and 20:00 (i.e., this is the interval in which the events occur), there is most traffic on Central Park Avenue. Note that the pattern occurs in about 24 hours. Other interesting patterns could be given in the context of social networks. For instance, the release of tweets from some specific places in a repetitive way at a particular moment of the week (e.g., tweets sent from a pub on Fridays between 22:00-24:00). On the other side, one of the foremost applications is given over the OLAP, MOLAP, and ROLAP databases since it is workable to realize an analysis of trends over time. To illustrate this, suppose that someone wants to analyze the sales by weeks; once done, it is necessary to analyze the sales by month. This latter is well-known as granularity, where the analysis can be conducted using time intervals such as days, weeks, and months, to mention a few.

As mentioned above, aiming to find the patterns, events or episodes that occur over time outlined in a specific sequence, considering the order in which the events take place. Later, an events' minimum threshold is given to determine patterns along with the length of the periods to be analyzed. Notably, in this paper, we addressed the search for cyclical patterns.

1.1 Problem Definition

Let $o(s', t)$ be a spatio-temporal object determined by a spatial location s' and a point in time t. An event e (i.e., a change in the object's location or shape) associated with an object o_x, considering a location s'_m at time t_i, is represented by $e(o_x, t_i)$. Without loss of generality, we assume that the space in which the objects can be located is partitioned into a set of $n \times n$ disjoint cells of equal size. Following this logic, s'_m corresponds to one of the cells in the set. A sequence of localized events denoted as S_x (i.e., for object o_x) is the set of events that takes place over a time series τ, such as $\tau = \{t_1, \ldots, t_n\}$, and $t_i < t_{i+1}$. An association rule for o_x is given as $X \rightarrow Y$, such that $X, Y \subseteq S_x$ and $X \cap Y \neq \emptyset$. The support of X, denoted as $sup(X)$, is the number of events in S_x that contain X.

Incidentally, each t_i can be seen as an hour, day, week, or any other interval of time, and so on. Using this rationale, if t_i represents a day, then seven consecutive t_i depict a week.

Definition 1: An association rule over the sequence S_x has a cycle $c(l, t_b, l_b)$, if the association rule appears every l^{th} unit of times starting at time t_b, considering an occurrence interval of length l_b, such that l is the length of the cycle (i.e., l can be seen as the number of periods), and whose support outperforms a threshold called sup_{min}.

Now consider a pattern denoted as X, such as $X \subseteq S_x$. We say X is an p-periodic pattern if the length of X is p; where p is called the period. A pattern occurs when a sequence of events meets the minimum support provided by the user. It should be noted that the user also provides the p value. For instance, Let $p = 3$ be over the sequence $S = \{\{e\}\{b\}\{c\}\{e\}\{d\}\{c\}\{e\}\{c\}\{c\}\}$. Accordingly, there are three possible subsequences with three events each in which the pattern $\{e\}\{*\}\{c\}$ (i.e.,$\{*\}$ can be any event) can appear. This means that the $sup(\{e\}\{*\}\{c\})$ is equal to 1 because $\{e\}\{*\}\{c\}$ appears in all subsequences (e.g., if the minimum support were 1/3, the pattern should occur at least once within the subsequences). Therefore, $\{e\}\{*\}\{c\}$ is a perfect periodic pattern (i.e., it appears exactly every three periods), while simultaneously being a partial pattern with period $p = 3$, as the second event is missing. Conversely, when a pattern is absent in the sequence, it is imperfect. Remarkably, a full periodic pattern or a partial pattern can be both perfect or imperfect.

Aiming to differentiate between cyclical patterns and periodic patterns, let us consider the following sequence: $\{a\}\{b\}\{c\}\{x\}\{y\}\{d\}\{e\}\{g\}\{x\}\{d\}$. When we set $p = 4$, it becomes infeasible to identify $\{x\}$ within any pattern, taking into consideration that a pattern must appear at least twice. However, if we establish a cycle composed of five periods (referred to as l, denoting the cycle length) and assume an interval of two events or periods, we can discern the subsequences $\{x\}\{y\}$ and $\{x\}\{d\}$. Hence, we can designate $\{x\}\{*\}$ as a partial cyclical pattern that repeats every five periods (where $l = 5$, $l_b = 2$, and $t_b = \{x\}$). In a broader context, it is important to note that a cyclical pattern aligns with a periodic pattern when $p = l = l_b$, and t_b can represent any event within the subsequence, even if the event reappears at the subsequent $+p$ position. Noteworthy, this latter form of cyclical pattern corresponds to a periodic cyclical pattern.

1.2 Contribution

The main contribution of this paper can be itemized as follows:

- A new algorithm called HashCycle, which is based on a hash table to detect cyclical patterns over spatio-temporal data. Different from other algorithms grounded on association rules HashCycle does not require minimum support.
- HashCycle's time complexity is presented in a context where time and space are considered on search patterns.

- A set of experiments that involve both synthetic and real data are exposed. The synthetic dataset is used to corroborate that the implement algorithms are sound (i.e., it means that they find the periodic cyclical patterns on the dataset), while the real dataset corresponds to a subset of issued Tweets from New York. Using the Tweets dataset was feasible to find cyclical patterns.
- Empirical results show that HashCycle is more efficient than PPA and Apriori algorithms in terms of processing times, which are in line with their respective time complexities.

The remainder of this paper is organized as follows. Section 2, reviews related work for detecting sequential patterns on spatio-temporal databases. In section 3, we present our proposed method for detecting cycle patterns. Section 5 describes empirical results. Finally, conclusions are presented in section 6.

2 Related Work

In the literature, methods for sequential pattern mining can be categorized into three groups: (i) methods that rely on mathematical approaches, (ii) methods that search for significant patterns using machine learning techniques, and (iii) methods that are based on association rules.

Group (i) encompasses algorithms that assess circular autocorrelation using Fourier transform [Malode et al., 2015, Parthasarathy et al., 2006]. Khanna and Kasurkar [Saneep and Swapnil, 2015] address three types of periodicity (symbol periodicity, segment periodicity, and partial periodicity) by proposing corresponding variants of autocorrelation-based algorithms. These methods are robust against noise and efficient in extracting partial periodic patterns with unknown periods and no additional domain knowledge. However, they cannot guarantee the extraction of all periodic patterns satisfying given conditions (e.g., minimum support). Other algorithms in this category have been proposed to find cycles in time series: the rainflow counting algorithm [Endo et al., 1974] and CyDeTS [Dambrowski et al., 2012, Gupta, 2022]. These algorithms detect cycles and provide cycle lengths that allow predicting life cycles for mechanical engineering applications and stationary electrical storages. Drawbacks include unsuitability for real-time applications, inefficient memory use, and the exclusion of spatial information.

Group (ii) includes machine learning approaches that require an objective function or a training dataset to define “good” sequential patterns [Jamshed et al., 2020, Bunker et al., 2021]. A primary drawback of these machine learning methods is their impracticality when users lack sufficient domain knowledge for managing training and tuning stages. We discard studying this category as it implies training and tuning stages.

Finally, group (iii) methods involve techniques primarily based on association rule mining algorithms. Many of these techniques derive from the ”Apriori” approach proposed by Agrawal and Srikant [Agrawal and Srikant, 1994]. Apriori explores frequent itemsets, pruning infrequent ones during the process to manage the number of item combinations, preventing explosion. While still popular [Tirumalasetty et al., 2015], applying Apriori to sequential pattern mining is impractical due to potentially examining 2^n candidates. Possible optimizations leverage inherent properties in periodic pattern mining [Ozden et al., 1998]. These authors propose two algorithms: one called the sequential algorithm, based on Apriori, and the second one called the interleaved algorithm. According to the authors, the improvement of the second approach with respect to the first one is about 5%. These algorithms explore unnecessary candidates and are not tested on real data but on randomly generated data. In this synthetic data, events are manipulated and taken to binary space for pruning, something that does not happen in the real world. For these reasons, these algorithms have been discarded from implementation and comparison in this paper.

It is relevant to highlight that the most related works [Agrawal and Srikant, 1994] and [Ozden et al., 1998] are not designed to work in a spatiotemporal context. For spatiotemporal data mining, several approaches have been developed for essential applications like disease diffusion analysis [Gao et al., 2018], user activity analysis [Lv et al., 2013], and local trend discovery in social media [Ishida, 2010, Cheng and Wicks, 2014]. Other approaches dealing with spatial information include encoding spatial features as discrete symbols (assigning symbols reflecting semantic regions), treating them as continuous variables [Pillai et al., 2012, Pillai et al., 2013], and formulating them as a dynamic graph mining problem [Lahiri and Berger-Wolf, 2010]. Koylu [Koylu, 2019] proposes a framework to model and visualize the semantic and spatiotemporal evolution of topics in interpersonal communication on Twitter. This work uses term frequencies to indicate that a topic has a strong relationship with a term. Here, a user sets the time period for pattern analysis. The framework produces a series of word-topic matrices that represent the prominent words and their probabilities in each topic and a series of document-topic matrices that represent the prominence of topics and their probabilities in each document. Then, these word-topic and document-topic matrices are used to group similar topics, allowing the identification of topic chains that illustrate temporally consistent topics. Finally, the topic evolution and spatiotemporal patterns are linked in a web-based geovisual environment. Gutiérrez et al.'s approach [Gutiérrez-Soto et al., 2022] aims to search periodic patterns, not cyclical ones. It computes the frequencies of events in a sequence and based on this factor discards some irrelevant patterns. It does not use any specific data structure (for example a hash table). Applying this approach to searching cyclical patterns would require mining and pruning all possible cycle periods. Moreover, this algorithm will not find cycles if they are not periodic.

An essential direction in this field involves optimizing approaches through improved algorithms and data structures. Han et al. [Han et al., 1999] propose the max-subpattern hit set algorithm, built upon a tailored data structure called max-subpattern tree, for efficiently generating larger partial periodic patterns. Yang et al. [Yang et al., 2013] propose a projection-based partial periodic patterns algorithm (PPA), more efficient than the max-subpattern hit set algorithm and Apriori for discovering partial periodic patterns.

To our knowledge, few works have addressed an efficient search for cyclical patterns over STDB and often neglect simultaneous time and space cost reduction. This paper introduces a novel and efficient algorithm (HashCycle) for identifying cyclical patterns over STBD. The discrete symbol encoding approach enables the exploitation of insights gained from mining sequential periodic patterns. HashCycle, inspired by a hash table, detects patterns through collisions. It doesn't demand minimum support, detecting all patterns appearing at least twice in a single execution. Our proposed algorithm demonstrates efficiency and scalability using real Twitter data.

3 Algorithms

In this section, an example has been given to clarify the steps involved in each algorithm, considering the temporal aspect.

3.1 Apriori Algorithm

Apriori checks all subsets and returns all those who achieved the minimum support. To achieve this goal, Apriori first acquires all frequent itemsets of size 1; after that, all frequent itemsets of size 2, and so on. Note that Apriori was not designed to consider the temporal aspect of events in a sequence. Therefore, considering the temporal aspect has a relevant impact on its performance. Aiming to show this impact, an example is given. Consequently, the example string can be perceived as a time series with period four, such that a letter represents an event. It is essential to highlight that the events in square brackets denote that these can occur in parallel (i.e., an event characterizes a particular object considering the space and temporal aspect).

$$a\left\{b,c,d\right\}eda\left\{b,c\right\}adabcbaccdaabc$$

According to the periodicity previously indicated, the number of periods in this series is four:

$$a\left\{b,c,d\right\}ed - a\left\{b,c\right\}ad - abcd - accd - aabc \tag{1}$$

The first step finds the L_1 set. L_1 involves all those candidates L_1 frequent that achieve the minimum support. Hereafter, F_k will be used instead of L_k . It is relevant to mention that F_k keeps the positions of the events in the sequence. From now, F_k will be employed in all algorithms. In the following example, we use the notation F_{yCand} to denote the frequency of candidate events in forming a pattern composed of y events Therefore, using minimum support of 3, we have:

$$\begin{aligned} F_{1Cand}: \quad & \{a***:\tfrac{5}{5}, *a**:\tfrac{1}{5}, **a*:\tfrac{1}{5} \\ & *b**:\tfrac{3}{5}, **b*:\tfrac{1}{5}, ***b:\tfrac{1}{5}, *c**:\tfrac{4}{5}, \\ & **c*:\tfrac{2}{5}, ***c:\tfrac{1}{5}, *d**:\tfrac{1}{5}, ***d:\tfrac{3}{5}\} \end{aligned}$$

$$F_1:\left\{a***:\frac{5}{5}, *b**:\frac{3}{5}, *c**:\frac{4}{5}, ***d:\frac{3}{5}\right\}$$

$$\begin{aligned} F_{2Cand}: \quad & \{ab**:\tfrac{3}{5}, ac**:\tfrac{3}{5}, a**d:\tfrac{3}{5}, \\ & *\left\{b,c\right\}**:\tfrac{2}{5}, *b*d:\tfrac{2}{5}, *c*d:\tfrac{3}{5}\} \end{aligned}$$

$$F_2:\left\{ab**:\frac{3}{5}, ac**:\frac{3}{5}, a**d:\frac{3}{5}, *c*d:\frac{3}{5}\right\}$$

Following the same steps gave by Apriori, the algorithm finishes when $F_K:\emptyset$

$$F_{3Cand}:\left\{a\left\{b,c\right\}**:No, ab*d:No, ac*d:\tfrac{3}{5}\right\}$$

$$F_3:\left\{ac*d:\frac{3}{5}\right\}$$

Following this thread, the number of events in F_3 does not generate a set of F_4 candidates. Hence, the algorithm finishes with F_3.

3.2 PPA Algorithm

The PPA algorithm works the following: first, it runs the sequence and divides it into l partial periods. Later, every event is codified considering its location within the partial period. This way, codified events are represented as a matrix, which the authors name as *EPSD*. In *EPSD*, the first row is determined by the first l codified events, such that each column represents the event position within the partial periods. Thus, it is possible to count the instances of each event by columns. This last step is to verify if events accomplish the required minimum support. To exemplify this, a particular instance of Eq(1) is employed, such as:

$$abed - abad - abcd - accd - aabc$$

$$\mathit{EPSD} = \begin{pmatrix} a_1 & b_2 & e_3 & d_4 \\ a_1 & b_2 & a_3 & d_4 \\ a_1 & b_2 & c_3 & d_4 \\ a_1 & c_2 & c_3 & d_4 \\ a_1 & a_2 & b_3 & c_4 \end{pmatrix}$$

where the element $x'_{j,i}$ in $EPSD$ corresponds to the event x in the position i in the sequence, such that the first subsequence ($j = 1$) is the first row in $EPSD$, the second subsequence ($j = 2$) corresponds to the second row, and so on.

A candidate subsequence is obtained once those events are counted by columns and achieve the minimum support. Subsequently, events that form this sub-sequence are sorted first, considering the partial positions and then considering each event's lexicographic nomenclature (the maximum nomenclature has size a). Note that the last sub-sequence S_c is equivalent to having F_1. According to the authors, S_c is used to look for the other F_{kCand} patterns. A sub-routine called *Finding-FTP* is used to achieve this goal, so each event of S_c is used as a prefix to obtain the patterns that comply with the minimum support over *EPSD*. Finally, all F_k that fulfil the minimum support are obtained.

3.3 HashCycle Algorithm

The underlying motivation for employing a Hash Table is to reduce processing times during insertion, particularly for pattern searching. Furthermore, the order in which events take place in the sequence (i.e., the possible periods) makes them easily located in the Hash Table slots. Once this is done (i.e., when the Hash Table contains all the events), patterns can be easily obtained by traversing the lists belonging to the slot where the events appear. The occurrence of these events maintains the order in the lists with a positional difference of p. It's important to mention that each slot can have multiple lists, where the first element in the list determines the appearance of the subsequent elements in that list. Indeed, the length of each list indicates whether the minimum support can be achieved. This last aspect makes the Hash Cycle extremely efficient when it comes to finding patterns. To achieve this goal, we utilize the modulo operator, which is widely used in Hash Tables, particularly when the data universe (i.e., keys) is numeric. The latter allows us to quickly locate the slot in which the event is. For example, calculating the modulus between $(4\ mod\ 2) = 0$ corresponds to the remainder of the division between 4 and 2. The previous one places us in slot 0, where the insertion or search takes place. It should be noted that computing the modulus operation takes $O(1)$ (see Line 4, where h is provided the modulus), which is very efficient. Notably, we use this operator in our algorithm. By the way, our algorithm's main disadvantage is the Table's size since the different periods imply different Tables. Nonetheless, this is not a disadvantage due to the memory capacity of conventional computers. The following paragraphs describe how our algorithm operates on the Hash Table.

HashCycle relies on a hash table; Table T contains m slots, which store the events; in turn, these m slots have other s slots (i.e., which can be at most p), each one such that each slot stores a list with candidate patterns, considering the position in which events appear in the sequence. In the beginning, HashCycle runs the sequence and checks if the event is in T, so whether the event is not in the table, the event is added in $T[e_j]$ (i.e., $1 \leq j \leq m$) considering at the same time the position in which the event occurs in the sequence. This latter determines the following slot s'_k, between 1 and p, storing the event in $T[e_j].s'_k$ (see line 4 of the Algorithm). On the contrary, if the event is in T, it is put in its corresponding list. Once the sequence has been run, it is workable to determine the patterns. To achieve this goal, the period has to be previously provided. Pattern cyclical are obtained from lists of each event; note that these are perfect cyclical patterns. F_1 is extracted by running all those lists having at least two elements. As regards F_2, this is built using the combination of F_1, i.e., combining all lists with two or more elements (many implementations can be found in several algorithms' books). Following this reasoning, F_k (it is achieved with Lines 2, 3, and 4 of the Algorithm) can be achieved in the same way. It should be noted that Algorithm 1 allows achieving all patterns, as it processes all slots of T along with their corresponding lists. Aiming to shed light suppose the following example given the following sequence $aac - aab - aa$. There are two lists in T (see Figure 1) with period three over a sequence formed by events a, b, and c, whose length is 8. A list contains the following events (a,1), (a, 4), (a, 7), meantime another list is composed of (a,2), (a,5), (a, 8), such that the first list can be seen as $T[a].s_1$, and the second one as $T[a].s_2$. On the other hand, suppose the following lists (a,1), (a, 4), (a, 7), and (b,2), (b,5), (b, 8), combining both is feasible to obtain the pattern (a, b,*).

Algorithm 1 HashCycle

```
Require: $sup_{min} \geq 0 \wedge S \neq \emptyset \wedge n \geq 0 \wedge m \geq 0 \wedge p \geq 0$
Ensure: $Patterns$
 1: $Patterns \leftarrow \emptyset$
 2: for $x \leftarrow 1, x \leq m$ do
 3:     for $i \leftarrow 1, i \leq n$ do
 4:         $T[e_i].s_{k'} \leftarrow h(p, e_i(o_x, t_i))$
 5:         if $T[e_i].n_{k'} \geq sup_{min}$ then
 6:             $T[e_i].n_{k'} \in Patterns$
 7:         end if
 8:     end for
 9:     return $Patterns$
10: end for
```

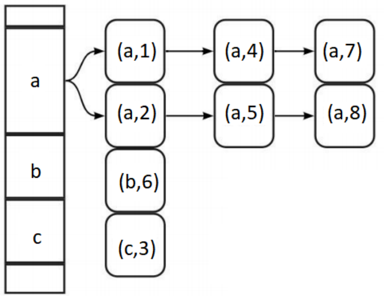


Figure 1: HashCycle's Data Structure

4 Time complexities

This section gives the time complexities for the algorithms Apriori, PPA, and HashCycle. Those time complexities are yielded in a Spatio-temporal context where the order in which the events occur plays a crucial role.

4.1 Apriori's time complexity

According to [Gutiérrez-Soto et al., 2022], Equation (2) represents all candidate patterns provided by Apriori such that m is the number of events, f indicates the size of a period, and m^f corresponds to the number of forms that can appear in the events. $C_{p,f}$ represent all ways to organize f events over a segment of size p. Owing to a pattern should appear at least twice over the sequence, and then it only is necessary to run the middle of the sequence; therefore, p can achieve $\frac{n}{2}$.

$$m^f * C_{p,f} = m^f * \binom{p}{f} = m^f * \frac{p!}{(p-f)! * f!} \tag{2}$$

The equation represents all candidate patterns for $f = 1$:

$$\sum_{p=2}^{n/2} m * p \tag{3}$$

Expression (3) can be implemented in any programming language using three loops. Recall that a pattern is considered as such if it appears at least twice within the sequence. In this manner, the outer loop runs from 1 to $\frac{n}{2}$ (i.e., $p = 2$ to $\frac{n}{2}$). This implies that half of the sequence could potentially form a pattern. The middle loop iterates through the sequence (i.e., it goes through n). These two loops provide $\frac{n^2}{2}$, which is equivalent to n^2. The innermost cycle runs from 1 to m (i.e., such that m are the events). Therefore, the time complexity of (3) is determined by:

$$O\left(mn^2\right) \tag{4}$$

Extending Equation (2), it is workable to obtain all candidate patterns (i.e., all candidates from F_{2Cand} to $F_{\frac{n}{2}Cand}$), thus:

$$\sum_{p=2}^{n/2}\sum_{F=1}^{p} m^F * \frac{m^F * p!}{(p-F)! * F!} \tag{5}$$

Note that the goal of Expression (5) is not to solve the expression itself (i.e., from a mathematical point of view), but instead to analyze how it can be implemented. It should be noted that many techniques can be used to calculate the factorial, such as dynamic programming. Due to this, the calculation is repeated many times. Therefore, the last one can be executed outside of both summations, making the use of each factorial a constant-time operation. On the other hand, it is evident that in the inner summation, $m^F * m^F$ appears, which from a programming perspective is essentially a multiplication by itself and operates in $O(1)$ time. Without loss of generality, expression (5) can be simplified as $n\sum_{F=1}^{n/2} m^F$, such that F takes the p value as maximum (see Expression (5)). From this latter expression, the n, which multiplies the summation, corresponds to a loop until n, while the summation itself is another cycle ranging from 1 to $\frac{n}{2}$. As a result, this last loop takes $O(m^{\frac{n}{2}})$. Therefore, the time complexity for Apriori is given by [Gutiérrez-Soto et al., 2022]:

$$O\left(n\sqrt{m^n}\right) \tag{6}$$

4.2 PPA's time complexity

Calculating the sub-sequence S_c is similar to computing F_1; Roughly, calculating the sub-sequence S_c is similar to computing F_1, Whereby it takes $O(mn^2)$. Additionally, two previous sorts are necessary before obtaining F_1. One must be computed for the partial position, while the second corresponds to the event's lexicographic nomenclature. The first sort takes $O(p \cdot log_2 p)$; meantime second implies $O(a \cdot \log_2 a)$, where p corresponds to the period, and a is the maximum nomenclature of an event. Both sorts can be carried out using MergeSort when p and a are remarkable. Lastly, a subroutine builds all prefixes in S_c (i.e., for each event) and verifies them inside *EPSD*. Finally, the time complexity for PPA is given by [30] (if $((plog_2 p) + (alog_2 a)) < (m^{m+1}n^2)$):

$$O(max((m^{m+1}n^2), (plog_2 p) + (alog_2 a))) = O(m^{m+1}n^2) \tag{7}$$

4.3 HashCycle's time complexity

Aiming to provide the time complexity for HashCycle, we introduce the following definitions:

Definition 2: Let $|S(\tau)|$ be a number of events over the sequence S.

Definition 3: Let $R\{e\}$ be the indicator random variable over a sequence S, which is associated with event e, such that:

$$R\{e(o_x, t_i)\} = \begin{cases} 1 & \text{if } e \text{ occurs at the } t_i \text{ moment over S.} \\ 0 & \text{if } e \text{ does not occur at the } t_i \text{ moment over S.} \end{cases}$$

Note that our sample space $\mathbb{S}$ is formed by probability (*H*), it is that e happens at the t_i moment over S, and the probability (*D*), where e does not happen at the t_i moment over S; thus $Pr\{H\} = Pr\{D\} = \frac{1}{2}$. Therefore, the indicator random variable $R\{H\}$ for $\mathbb{X}_H$ can be written as: $\mathbb{X}_H = R\{H\}$.

Lemma 1:
Given $\mathbb{S}$ and an event $e(o_x, t_i)$ over S, let $\mathbb{X}_e = R\{e\}$.
Then $E[\mathbb{X}_e] = Pr\{e\}$.
Proof:
By the *Definition 3* and the definition of expected value:

$$\begin{aligned} E[\mathbb{X}_e] &= E[R\{e\}] \\ &= 1 \cdot Pr\{e\} + 0 \cdot Pr\{\overline{e}\} \\ &= Pr\{e\}. \end{aligned}$$

such that $Pr\{\overline{e}\}$ corresponds to the complement of e. ■

Definition 4: Let α be the load factor for a hash table T, such that $\alpha = \frac{n}{m'}$ corresponds to the average number of elements stored in a chain in T. T has m' slots that stores $|S(\tau)| = n$ occurrences of events. In our particular, case $m' = mp$ where m are events, and p is the period. In simple words, each event e_j has s slots, which are between 1 and p. The events are stored in T along with their position that appears in $S(\tau)$.

Definition 5: Let h the hash function over the hash table T, such that:
$h(p, e_j(o_x, t_i)) \to T[e_j].s_{k'}$, where $k' = (i \ mod \ p)$, and mod provides modulus or remainder of a division. Note that i determines the position of an event inside $S(\tau)$.

Definition 5: Let $n_{k'}$ the length of the list $T[e_j].s_{k'}$, so that n is

$$\sum_{j=1}^{m} \sum_{k'=1}^{p} T[e_j].n_{k'}$$

Theorem 1:
In T, an unsuccessful search takes $\Theta(1 + \alpha)$, under the assumption of simple uniform hashing.
Proof:
An event e_j that does not store yet in the table is equally likely to hash to any of the m' slots, assuming a simple uniform hashing. The expected time for a search unsuccessfully implies achieving the end of the list $T[h(p, e_j(o_x, t_i))].n_{k'}$, which has expected length $E[n_{h(p,e_j(o_x,t_i))}] = \alpha$. Therefore, the expected number of events reviewed in an unsuccessful search is α, and the total time required is $\Theta(1 + \alpha)$.■

Theorem 2:
In T, a successful search can be solved in $\Theta(1 + \alpha)$, under the assumption of simple uniform hashing.
Proof:
Let e_j denote the event to be searched on T. Without loss of generality, it is workable to have the list from $T[e_j].s_{k'}$ (by Definition 5). As the distribution is uniform and the expected length of $T[e_j].n_{k'}$ corresponds to $E[n_{h(p,e_j(o_x,t_i))}] = \alpha$ (by Definition

4), then the search algorithm performs no more extended than $1 + \alpha$ comparisons (by Theorem 1). As a consequence, a successful search can be solved in $\Theta(1 + \alpha)$.■.

Consider that the time required to search an event depends on the length $n_{k'}$ of the list $T[e_j].s_{k'}$. Towards that end, the hash function $h(p, e_j(o_x, t_i))$ is computed in $O(1)$.

Incidentally, on this occasion, we have used the Θ notation since the worst-case is given when all the events are in the same list, by which, it should take $O(n)$. Nevertheless, we assume that the events follow a uniform probabilistic distribution; hence, the lists in the table are likely to have similar lengths. In this way, the Θ notation is more appropriate for this analysis.

5 Experiments

5.1 Experimental Environment and Empirical Results

Our experiments were carried out over two data types: synthetic and actual. Actual data corresponds to tweets. The algorithms were implemented in Java. To obtain correct measurements of processing times on JVM (Java Virtual Machine), JMH (Java Microbenchmark Harness) has been used. JMH is a Java framework used to build, run, and analyze benchmarks written in Java and other JVM languages. By doing this, it is possible to isolate the algorithms' processing times, ignoring elements such as caching, pipelines, and processing priority. JMH included ten training iterations and fifteen measurement iterations in all the experiments.

All the results are shown in average times (Avg)— times are expressed in milliseconds in Tables I-IV — along with its standard deviation (SDev), the latter to show the variability of the algorithms regarding the data. All experiments were executed in an Intel Core i3-5005U 2GHz; 8GB 1600 MHz DDR3L; and Windows Operating System 10 Home 64 bits.

Experimentation over synthetic data aims to check that all algorithms are sound (i.e., all algorithms find and return the same patterns). Two events' sequences have been created, such that the events occur randomly over the sequences (note that the events are known, and the uniform distribution has been used to allow them to occur in different places within the sequence. The first sequence does not contain patterns (see Table I). In the second sequence, a pattern of length ten has been added (see Table II). In both sequences, the experiments involve searching for patterns of lengths 10, 50, and 100 on sequences of lengths 500 and 1000.

Tables I and II show the average running times over a synthetic dataset. HashCycle presents the best performance – in terms of processing times — considering all scenarios (i.e., lengths, patterns, and sequences). Furthermore, HashCycle's standard deviations are the lowest, indicating that its behaviour is scalable independent of the pattern's length to search and the sequence's length. PPA achieves the second-best performance; nonetheless, it presents a higher standard deviation than HashCycle. Finally, the worst results are given by Apriori, where in some cases it outperforms 20000 milliseconds by which its results are omitted (the results appear as "-' ' in Table II).

Regarding real data, these involve four New York districts; Brooklyn, The Bronx, Manhattan, and Queens. Aiming to homogenize the number of inhabitants, we have built two groups: The Bronx - Brooklyn (designated as the L1 sector) and Manhattan - Queens (called the L2 sector). Incidentally, using the endpoints of a Twitter API, it is possible to identify the tweets issued from both sectors. The schedule used in the analysis of the tweets corresponds to prime time, which is between 21:00 and 21:30. To identify time ranges in which tweets are emitted, three periods, $\mathbb{T}_1$=21:00-21:10, $\mathbb{T}_2$=21:10-21:20, and $\mathbb{T}_3$=21:20-21:30, have been defined. Using sectors and periods, we can define a set of events; $\mathbb{E} = \{A, B, C, D, E, F\}$. It should be noted that the events associated with the first period ($\mathbb{T}_1$) only include events A and B, while the events linked to the second period ($\mathbb{T}_2$) are only composed of events C and D; the same reasoning applies to the third period. The data collected comprised a week, starting on January 27th, 2021. Therefore, the sequences used by the three algorithms involve all the week's events. In this way, there are three sequences $S_{\mathbb{T}_1}$, $S_{\mathbb{T}_2}$, and $S_{\mathbb{T}_1}$ (i.e., the sequences associated with the periods $\mathbb{T}_1$, $\mathbb{T}_2$, and $\mathbb{T}_3$, respectively) (see columns in Table 2). Finally, it is worth mentioning that in these experiments, lengths between 2 and 100 have been used to search for patterns, considering that all $S_{\mathbb{T}_i}$ comprise more than 1000 event occurrences each.

Tables III and IV display the average running times with actual data extracted from Twitter. Following the trending of Tables I and II, HashCycle presents the best processing times in all scenarios. Once again, HashCycle's standard deviations are the lowest. On the other hand, the second-best performances are accomplished by PPA; similarly to Tables I and II, PPA has higher standard deviations than HashCycle. Apriori provides the worst results.

In summary, HashCycle presents the best performances and provides the lowest standard deviations. It is worth stressing that running times present a significant saving regarding its adversaries. Finally, the second-best performance is achieved by PPA, followed by Apriori.

		Synthetic Without Patterns			
		500		1000	
Algorithm	Length	Avg	SDev	Avg	SDev
Apriori	10	0,790	0,009	1,529	0,009
Apriori	50	9,154	0,099	15,231	0,139
Apriori	100	35,727	0,329	55,294	0,424
PPA	10	0,249	0,003	0,523	0,006
PPA	50	0,547	0,007	1,637	0,009
PPA	100	0,791	0,005	2,372	0,016
HashCycle	10	0,017	0,001	0,038	0,001
HasgCycle	50	0,031	0,001	0,065	0,001
HashCycle	100	0,047	0,001	0,099	0,002

Table 1: Experiments with synthetic data without patterns.

		Synthetic With Patterns			
		500		1000	
Algorithm	Length	Avg	SDev	Avg	SDev
Apriori	10	0,667	0,049	1,264	0,013
Apriori	50	1779,36	14,402	2548,39	63,68
Apriori	100	-	-	-	-
PPA	10	0,232	0,011	0,468	0,007
PPA	50	3,520	0,051	6,894	0,131
PPA	100	3745,8	161,01	6532,43	192,94
HashCycle	10	0,019	0,001	0,044	0,001
HasgCycle	50	0,043	0,001	0,068	0,001
HashCycle	100	0,071	0,001	0,144	0,009

Table 2: Experiments with synthetic data with patterns.

		S_{T_1}		S_{T_2}	
Algorithm	Length	Avg	SDev	Avg	SDev
Apriori	10	0,952	0,007	1,197	0,012
Apriori	50	9,822	1,500	10,888	0,332
Apriori	100	28,601	0,190	42,312	0,408
PPA	10	0,366	0,010	0,465	0,005
PPA	50	0,633	0,010	0,919	0,010
PPA	100	0,865	0,010	1,178	0,010
HashCycle	10	0,044	0,001	0,060	0,001
HasgCycle	50	0,113	0,008	0,134	0,005
HashCycle	100	0,171	0,002	0,224	0,008

Table 3: Experiments with data extracted from Twitter (S_{T_1} and S_{T_2}).

6 Discussion

As mentioned earlier, a cyclical pattern is equivalent to a periodic pattern if the cyclical pattern can be expressed as a periodic pattern (i.e., it is a periodic cyclical pattern). Based on this foundation, it is feasible to detect periodic cyclical patterns using those algorithms to discover periodic patterns. Achieving this goal implies checking all possible periods to detect a cycle (i.e., this involves examining all possible periods from 1 to $\frac{n}{2}$, such as n is the sequence's length).

		S_{T_3}	
Algorithm	Length	Avg	SDev
Apriori	10	1,245	0,015
Apriori	50	13,446	0,918
Apriori	100	36,346	0,327
PPA	10	0,517	0,014
PPA	50	0,930	0,010
PPA	100	1,207	0,007
HashCycle	10	0,060	0,001
HasgCycle	50	0,146	0,001
HashCycle	100	0,236	0,002

Table 4: Experiments with data extracted from Twitter (S_{T_3}).

On the other hand, a few algorithms to look for cyclical patterns can be found in the literature. Moreover, some of them use mathematical techniques such as autocorrelation, which can not be used to compare with our algorithm. Instead, we have chosen the PPA algorithm, which has an excellent performance. Although PPA has been designed to find perfect and imperfect period patterns, it can find periodic cyclical patterns. Similarly, we have used the Apriori algorithm to have an upper bound since its performance has exponential behaviour. One should observe that both PPA and Apriori use minimum support, which is working to better compare the processing times among them

As already mentioned, a cyclical pattern is equivalent to a periodic pattern; however, the converse is not true. Furthermore, let's consider the following sequence $S = \{a\}\{f\}\{g\}\{x\}\{d\}\{c\}\{e\}\{c\}\{x\}\{c\}\{c\}\{e\}\{c\}\{h\}\{x\}$. If we consider $l = 5$, and $l_b = 2$, with $t_b = \{a\}$, then $\{\{x\}\{d\}, \{x\}\{c\}, \{h\}\{x\}\}$ are possible patterns, however, $\{x\}$ is always within $\{+\}\{+\}$ with $l_b = 2$ (i.e., $\{x\}$ can be in any of the two positions within the l_b). We denominated $\{+\}$, when an event such as $\{x\}$ occurs in any of all position with $\{+\}$. The latter provides new patterns that can occur within an interval, but only sometimes in the same position. For example, consider a postman who usually has to deliver parcels on an island, sometimes on Monday or Tuesday, depending on the weather. In this way, Monday and Tuesday do not correspond to a period from a periodic point of view; this is in contrast to a cyclical pattern, where both days conform to a recognizable pattern. In that regard, please note that in this paper, we are only concerned with proving the algorithms' efficiency when the cyclical patterns are periodic. Nonetheless, we believe that looking for cyclical patterns that there are not periodic is a good challenge since this type of problem is not common in Data Science.

On the other side, in this paper, we did not analyze the tweets' content but rather analyzed the time and place from where those are emitted. These cyclical patterns can indicate some interaction among users, suggesting something is happening around them. Furthermore, we believe that finding this kind of pattern makes it possible to analyze tweets' content better. In this way, spatio-temporal patterns can be exceptionally useful in social networks since these can prune tweets whose content is irrelevant. This latter can provide more efficient processing of tweets' content.

Regarding running times, the minimum support was instantiated in 1 in all experiments. Roughly, the running times decrease considerably when the minimum support increases, even for the Apriori algorithm. This latter is due to a significant pruning occurring every time the minimum support is raised. On the other hand, a successful search takes a maximum time $\Theta(1+\alpha)$ when it achieves the list's end (according to *Theorem 2*). Accordingly, further experimentations should be conducted by varying the length of the interval (i.e. l_b) in which the patterns can occur.

7 Conclusion and Future Work

This paper introduces a novel and efficient algorithm named HashCycle, designed to identify cyclical patterns on Spatio-temporal data. HashCycle is characterized as a hash Table, where individual slots correspond to the events within a given sequence. Each slot contains linked lists that represent the relative positions of events, with these positions having a maximum length equal to the cycle period. Linked lists are obtained through collisions, ensuring that only those lists containing two or more nodes contribute to the formation of a valid cyclical pattern. Notably, HashCycle demonstrates exceptional space-saving capabilities by discarding lists that cannot form a complete cycle. Additionally, it is essential to highlight that HashCycle operates effectively without requiring a minimum support threshold, enabling the detection of all cyclical patterns appearing at least twice within a single execution. The time complexity for HashCycle implies $\Theta(1+\alpha)$, while PPA takes $O(m^{m+1}n^2)$, and Apriori achieves $O\left(n\sqrt{m^n}\right)$. The algorithms were evaluated over two Spatio-temporal datasets, one synthetic and another real, considering in this latter the location where tweets were emitted. Empirical results show that HashCycle is not only more efficient than Apriori and PPA but also scalable.

Ideas for future work include two threads. The first thread involves implementing algorithms such as MS-Apriori, FP-Growth, and Max-Subpattern; extending the experimental environments; increasing periods and sequence lengths; and considering different supports. The second thread involves varying the length of the interval within which the patterns can occur. The latter should yield new patterns that are discarded when the patterns are periodic.

Acknowledgements

This research was supported by Universidad del Bío-Bío, Chile, under Grants No. DIUBB GI 195212/EF, INN I+D 23-53, and DIUBB 2130253 IF/R.

References

[Agrawal and Srikant, 1994] Agrawal, R. and Srikant, R. (1994). Fast algorithms for mining association rules in large databases. In *Proceedings of the 20th International Conference on Very Large Data Bases*, VLDB '94, page 487–499, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

[Bhattacharya et al., 2022] Bhattacharya, S., Somayaji, S. R. K., Gadekallu, T. R., Alazab, M., and Maddikunta, P. K. R. (2022). A review on deep learning for future smart cities. *Internet Technology Letters*, 5(1):e187.

[Bora et al., 2013] Bora, Zaytsev, Chang, and Maheswaran (2013). Gang networks, neighborhoods and holidays: Spatiotemporal patterns in social media. In *2013 International Conference on Social Computing (SocialCom)*, volume 0, pages 93–101.

[Bunker et al., 2021] Bunker, R., Fujii, K., Hanada, H., and Takeuchi, I. (2021). Supervised sequential pattern mining of event sequences in sport to identify important patterns of play: An application to rugby union. *PLoS One*, 16(9):e0256329.

[Cheng and Wicks, 2014] Cheng, T. and Wicks, T. (2014). Event detection using twitter: a spatio-temporal approach. *PLoS One*, 9(6):e97807.

[Dambrowski et al., 2012] Dambrowski, J., Pichlmaier, S., and Jossen, A. (2012). Mathematical methods for classification of state-of-charge time series for cycle lifetime prediction. In *Advanced Automotive Battery Conference*, Europe.

[Endo et al., 1974] Endo, T., Mitsunaga, K., Takahashi, K., Kobayashi, K., and Matsuishi, M. (1974). Damage evaluation of metals for random or varying loading—three aspects of rain flow method. *Mechanical Behavior of Materials*, (1):371–380.

[Gaede and Günther, 1998] Gaede, V. and Günther, O. (1998). Multidimensional access methods. *ACM Comput. Surv.*, 30(2):170–231.

[Gao et al., 2018] Gao, Y., Wang, S., Padmanabhan, A., Yin, J., and Cao, G. (2018). Mapping spatiotemporal patterns of events using social media: a case study of influenza trends. *Int. J. Geogr. Inf. Sci.*, 32(3):425–449.

[Gupta, 2022] Gupta, M. (2022). Mathematically determining patterns in time series with codes. http://bit.ly/490WJrU.

[Gutiérrez-Soto et al., 2022] Gutiérrez-Soto, C., Gutiérrez-Bunster, T., and Fuentes, G. (2022). A new and efficient algorithm to look for periodic patterns on spatio-temporal databases. *J. Intell. Fuzzy Syst.*, 42(5):4563–4572.

[Han et al., 1999] Han, J., Dong, G., and Yin, Y. (1999). Efficient mining of partial periodic patterns in time series database. In *Proceedings 15th International Conference on Data Engineering*, pages 106–115.

[Ishida, 2010] Ishida, K. (2010). Periodic topic mining from massive amounts of data. In *2010 International Conference on Technologies and Applications of Artificial Intelligence*, pages 379–386.

[Jamshed et al., 2020] Jamshed, A., Mallick, B., and Kumar, P. (2020). Deep learning-based sequential pattern mining for progressive database. *Soft Computing*, 24(22):17233–17246.

[Kandt and Batty, 2021] Kandt, J. and Batty, M. (2021). Smart cities, big data and urban policy: Towards urban analytics for the long run. *Cities*, 109:102992.

[Koylu, 2019] Koylu, C. (2019). Modeling and visualizing semantic and spatio-temporal evolution of topics in interpersonal communication on twitter. *International Journal of Geographical Information Science*, 33(4):805–832.

[Lahiri and Berger-Wolf, 2010] Lahiri, M. and Berger-Wolf, T. Y. (2010). Periodic subgraph mining in dynamic networks. *Knowl. Inf. Syst.*, 24(3):467–497.

[Li et al., 2019] Li, X., Chen, W., Chan, C., Li, B., and Song, X. (2019). Multi-sensor fusion methodology for enhanced land vehicle positioning. *Inf. Fusion*, 46:51–62.

[Li et al., 2011] Li, Z., Han, J., Ji, M., Tang, L.-A., Yu, Y., Ding, B., Lee, J.-G., and Kays, R. (2011). MoveMine: Mining moving object data for discovery of animal movement patterns. *ACM Trans. Intell. Syst. Technol.*, 2(4):1–32.

[Lü et al., 2019] Lü, G., Batty, M., Strobl, J., Lin, H., Zhu, A.-X., and Chen, M. (2019). Reflections and speculations on the progress in geographic information systems (GIS): a geographic perspective. *Int. J. Geogr. Inf. Sci.*, 33(2):346–367.

[Lv et al., 2013] Lv, M., Chen, L., and Chen, G. (2013). Mining user similarity based on routine activities. *Inf. Sci.*, 236:17–32.

[Malode et al., 2015] Malode, Y. B., Khadse, D. B., and Jamthe, D. V. (2015). Efficient periodicity mining using circular autocorrelation in time series data. *International Research Journal of Engineering and Technology (IRJET)*, 03(3):430–436.

[Molinaro and Orzes, 2022] Molinaro, M. and Orzes, G. (2022). From forest to finished products: The contribution of industry 4.0 technologies to the wood sector. *Computers in Industry*, 138:103637.

[Nakata and Takeuchi, 2004] Nakata, T. and Takeuchi, J.-I. (2004). Mining traffic data from probe-car system for travel time prediction. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '04, pages 817–822, New York, NY, USA. Association for Computing Machinery.

[Ozden et al., 1998] Ozden, B., Ramaswamy, S., and Silberschatz, A. (1998). Cyclic association rules. In *Proceedings 14th International Conference on Data Engineering*, pages 412–421.

[Parthasarathy et al., 2006] Parthasarathy, S., Mehta, S., and Srinivasan, S. (2006). Robust periodicity detection algorithms. In *Proceedings of the 15th ACM international conference on Information and knowledge management*, CIKM '06, pages 874–875, New York, NY, USA. Association for Computing Machinery.

[Pillai et al., 2013] Pillai, K. G., Angryk, R. A., and Aydin, B. (2013). A filter-and-refine approach to mine spatiotemporal co-occurrences. In *Proceedings of the 21st ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, SIGSPATIAL'13, pages 104–113, New York, NY, USA. Association for Computing Machinery.

[Pillai et al., 2012] Pillai, K. G., Angryk, R. A., Banda, J. M., Schuh, M. A., and Wylie, T. (2012). Spatio-temporal co-occurrence pattern mining in data sets with evolving regions. In *Proceedings of the 2012 IEEE 12th International Conference on Data Mining Workshops*, ICDMW '12, pages 805–812, USA. IEEE Computer Society.

[Rejeb et al., 2022] Rejeb, A., Suhaiza, Z., Rejeb, K., Seuring, S., and Treiblmaier, H. (2022). The internet of things and the circular economy: A systematic literature review and research agenda. *Journal of Cleaner Production*, 350:131439.

[Revesz, 2009] Revesz, P. (2009). *Spatiotemporal Interpolation Algorithms*, pages 2736–2739. Springer US, Boston, MA.

[Saneep and Swapnil, 2015] Saneep, K. and Swapnil, K. (2015). Design & implementation of efficient periodicity mining technique for time series data. *International Journal of Advanced Research in Computer and Communication Engineering*, 4(5):439–444.

[Song et al., 2015] Song, Y., Hu, Z., Leng, X., Tian, H., Yang, K., and Ke, X. (2015). Friendship influence on mobile behavior of location based social network users. *Journal of Communications and Networks*, 17(2):126–132.

[Tirumalasetty et al., 2015] Tirumalasetty, S., Jadda, A., and Edara, S. R. (2015). An enhanced apriori algorithm for discovering frequent patterns with optimal number of scans.

[Yang et al., 2013] Yang, K.-J., Hong, T.-P., Chen, Y.-M., and Lan, G.-C. (2013). Projection-based partial periodic pattern mining for event sequences. *Expert Syst. Appl.*, 40(10):4232–4240.

DOI: 10.17185/duepublico/81675
URN: urn:nbn:de:hbz:465-20240228-090549-3

JUCS - Journal of Universal Computer Science, 2023, 29(11): Special Issue.
https://lib.jucs.org/issue/4670