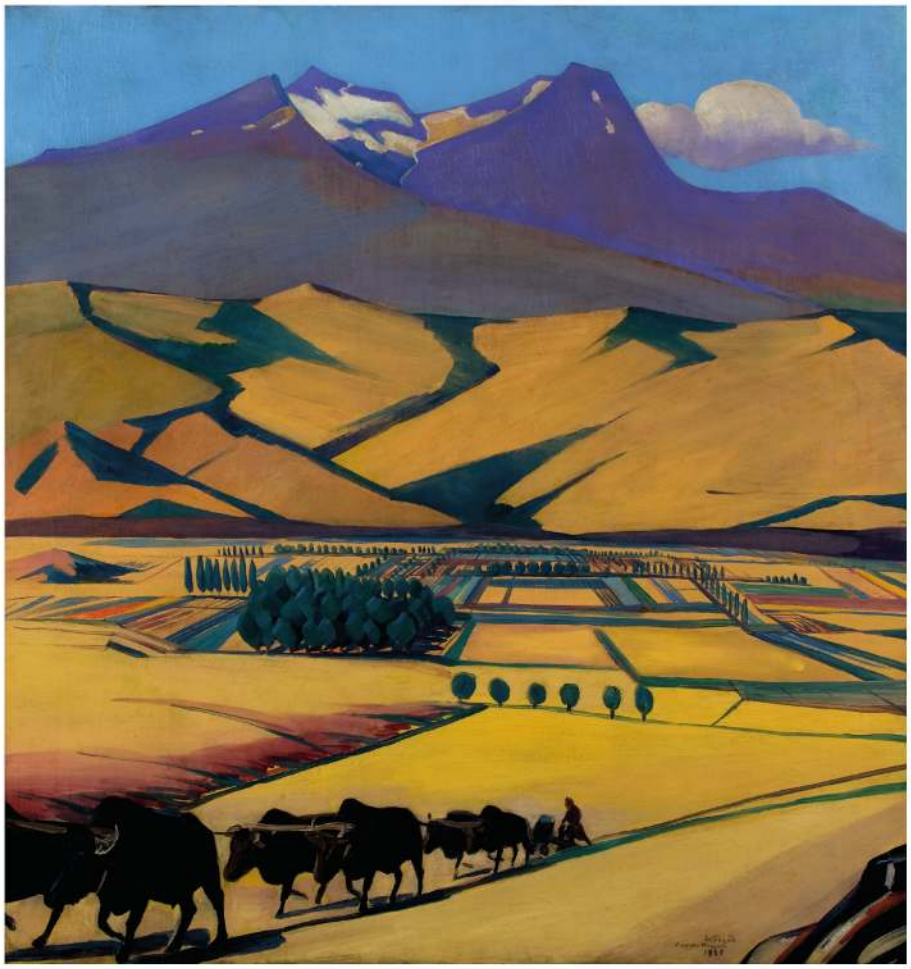
Data Science and Reliable Machine Learning

A.Hajian, N. Baloian, T. Inoue, W. Luther (Eds.)

Proceedings from the 4th Codassca Workshop
Yerevan, Armenia, October 2024



λογος

Aram Hajian, Nelson Baloian, Tomoo Inoue, Wolfram Luther (Eds.)

Data Science and Reliable Machine Learning

Logos Verlag Berlin

Bibliographic information published by Die Deutsche Bibliothek

Die Deutsche Bibliothek lists this publication in the Deutsche Nationalbibliografie; detailed bibliographic data is available in the Internet at http://dnb.d-nb.de.

The Open Access publication of this title was made possible with the support from the Publication Fund of the University Library of Duisburg-Essen.

Logos Verlag Berlin GmbH 2024

ISBN 978-3-8325-5855-0

DOI 10.30819/5855

Logos Verlag Berlin GmbH
Georg-Knorr-Str. 4, Geb. 10,
12681 Berlin
Tel.: +49 (0)30 / 42 85 10 90
Fax: +49 (0)30 / 42 85 10 92
https://www.logos-verlag.com

Aram Hajian, Nelson Baloian, Tomoo Inoue, Wolfram Luther (Eds.)

Data Science and Reliable Machine Learning

4th International Workshop at the American University of Armenia, College of Science & Engineering

October 3–6, 2024

Co-organized with IEEE Computer Society Armenia Chapter and

Supported by the Foundation for Armenian Science and Technology

Revised contributions

Volume Editors

Aram Hajian
American University of Armenia, College of Science and Engineering
40 Marshal Baghramyan Ave, Yerevan 0019, Armenia
Email: ahajian@aua.am

Nelson Baloian
Department of Computer Science, Universidad de Chile
Blanco Encalada 2120, Santiago 6511224, Chile
E-mail: nbaloian@dcc.uchile.cl

Tomoo Inoue
University of Tsukuba, Faculty of Library, Information and Media Science
1-2, Kasuga, Tsukuba, Ibaraki, 305-8550, Japan
E-mail: inoue@slis.tsukuba.ac.jp

Wolfram Luther
University of Duisburg-Essen, Scientific Computing, Computer Graphics, and Image Processing
Lotharstr. 65 (LF), 47057 Duisburg, Germany
E-mail: wolfram.luther@uni-due.de

2012 ACM CCS: Security and privacy · Computer system organization · Computing methodologies · Applied computing · Human-centered computing

PREFACE

After three successful versions of the Workshop on Collaborative Technologies and Data Science in Smart City Applications (CODASSCA 2018, 2020, and 2022) we are glad to present proceedings of the fourth version held at the American University of Armenia (AUA), October 3–6, 2024 in Yerevan, Armenia.

Data Science and Reliable Machine Learning (ML) presents seventeen selected and carefully revised papers, which are available at the start of the workshop in the Open Access Proceedings and as a printed version.

Society, technologies, and sciences are undergoing a rapid and revolutionary shift towards the integration of artificial intelligence into every system that people use in everyday life to create smart environments (SmE) through ambient intelligence (AmI) in highly connected and collaborative scenarios. The main source and asset for making smart systems is data, produced today in extraordinary large quantities. Data volumes are growing rapidly due to environmental sensor reading, sensor networks, broadband services, multimodal communication, whereas pervasive and embedded computing is enhancing the capability of everyday objects and easing collaboration among people.

Data from all areas of daily life that are increasingly accessible to a broad public enable the conception, creation, calibration, and validation of a process or complex systems. However, this requires international standards for data quality and access.

Mobile systems could enhance the possibilities available for designers and practitioners. Effective analysis, quality assessment and utilization of big data is a key factor for success in many business and service domains, including the domain of smart systems. Major industrial domains are on the way to perform this tectonic shift based on Big Data, Artificial Intelligence, Collaborative Technologies, Smart Environments (SmE) supporting Virtual and Mixed Reality Applications, Multimodal Interaction and Reliable Visual and Cognitive Analytics.

However, many requirements must be fulfilled and complexities resolved before we can effectively and efficiently turn the huge amount of generated data into information and knowledge. The first one is to ensure data quality, which includes accuracy and integrity of the obtained data, timely delivery, suitable quantity, etc. Privacy and security requirements and thorough end-to-end rights complement realization and deployment of modern design, software development and evaluation tools. The second one is to create understandable models, which can turn data into valuable information and then into knowledge.

A general challenge is the low interpretability of various artificial intelligence (AI) approaches and ML models, for which it is important that data scientists design the models, users understand results and developers debug and improve the models. The increasing complexity, limited explainability, and interpretability of the complex ML models make it difficult to address the emerging requirements for acceptance of these models and hinders their applications in industrial and mission critical scenarios.

Therefore, explainability, interpretability, transparency, and accountability of ML models and systems need to be further developed for an effective use of AI technologies. They are a prerequisite for a reliable application of AI within many problem areas, e.g., natural language processing, risk prediction in healthcare, fault/anomaly detection, computer vision or classification and regression under uncertainty, which are significant ML tasks. Researchers and practitioners working on theoretical and practical aspects of data science and reliable machine learning, as well as related and fundamental topics of statistical analysis, information transfer and processing were invited to attend this workshop.

The volume consists of the front matter and three thematic sections with seventeen peer-reviewed contributions.

The editors would like to express their gratitude to the Foundation for Armenian Science and Technology, the German Research Foundation (DFG) and the German Academic Exchange Service (DAAD) for funding their activities; to Yanling Chen, Rubina Danilova, Amalya Hambardzumyan, Ashot Harutyanyan, and Gregor Schiele for their ongoing encouragement and support, our reviewers and to all participants for their presentations and contributions to the workshop and this proceedings volume. We are particularly grateful to the director of the Martiros Sarian House–Museum in Yerevan, Mrs. Rouzan Sarian, for allowing us to use her grandfather's painting of Aragats as the cover picture for this volume.

Yerevan, Tsukuba, Duisburg, October 2024

The Editors: Aram Hajian, Nelson Baloian, Tomoo Inoue, and Wolfram Luther

TECHNICAL COMMITTEES

Organizing Committee Chair: Aram Hajian and Amalya Hambardzumyan (Armenia), Nelson Baloian (Armenia, Chile), Gregor Schiele (Germany)

Program Committee Chair: Ashot Harutyunyan (Armenia), José A. Pino (Chile), Wolfram Luther (Germany), Tomoo Inoue (Japan)

Track Data Science and Information Theoretic Approaches for Smart Systems: Yanling Chen, Han Vinck (Germany)

Track Collaborative Technologies with Applications in Smart Cities: Nelson Baloian, José Pino (Chile)

Track Smart Human-Centered Computing: Tomoo Inoue (Japan), Wolfram Luther (Germany)

Track Artificial Intelligence, Neural Networks and Deep Learning: Ashot Harutyunyan (Armenia), Gregor Schiele (Germany)

Track Technical Challenges for Smart Environments, Large Language Models: Gregor Schiele (Germany), Ashot Harutyunyan (Armenia)

CONTENTS

Outer Bound for Rate-Reliability-Equivocation Region of Compound Wiretap Channel with Informed Terminals

Mariam Haroutunian[1] [0000−0002−9262−4173]

Institute for Informatics and Automation Problems,
National Academy of Sciences of Armenia, Yerevan, Armenia
armar@sci.am

Abstract. The goal in designing communication systems in the presence of a wiretapper is to ensure that the message remains confidential between the transmitter and the intended receiver while minimizing the information available to the eavesdropper. Here we investigate the compound wiretap channel model, which is the extension of the wiretap channel, when the channels to the legitimate receiver and to the wiretapper depends on the number of possible states. Various cases can be considered, when these states are known or unknown to legitimate terminals.
We investigate the E-capacity-equivocation region which is the closure of the set of all achievable rate-reliability and equivocation pairs, where the rate–reliability function represents the optimal dependence of rate on the error probability exponent (reliability). Here the outer bound of this region is constructed in the case, when the states of the main channel are known to the legitimate terminals. A similar result for the case with unknown states was published previously.

Keywords: Compound Wiretap Channel · Channel Capacity · Rate-Reliability-Equivocation Region.

1 Introduction

A **wiretap channel**, in the context of communication systems and information theory, refers to a communication channel that is tapped or intercepted by a third party with the intention of eavesdropping on the communication. This concept is often used in the study of secure communication and cryptography.

In a wiretap channel model, there are typically three parties involved [1]:

Transmitter: This is the source that is trying to send a message to a legitimate receiver.

Receiver: The intended recipient of the message.

Eavesdropper: A third party that is intercepting or tapping into the communication channel in an attempt to gain unauthorized access to the message.

The transmitter wishes to send a message m to the receiver while keeping it as secret as possible from the eavesdropper. The information-theoretic investigation of generalized model of wiretap channel can be found in [2] - [6]. The wiretap

channel model is fundamental in understanding and designing secure communication systems, particularly in scenarios where privacy and confidentiality are of utmost importance.

The discrete memoryless **compound channel** is the model, when the channel depends on parameter $s \in \mathcal{S}$ and is invariable during transmission of one codeword of length N, but can be changed arbitrarily for transmission of the next codeword. This system can be considered in four cases, when the current state s of the channel is known or unknown at the encoder and at the decoder. This model was introduced and studied by Wolfowitz [7], who has shown that the knowledge of the state s at the decoder does not improve the asymptotic characteristics of the channel. So it is enough to study the channel in two cases, when the state of the channel is known or unknown at the encoder and decoder. This channel was investigated also in [8].

The **compound wiretap channel** introduces further complexity and considerations. In this model the channels to the legitimate receiver and to the wiretapper depends on the number of possible states. Similar to compound channel it is enough to study two cases:

- Case 1. The states of the channel are unknown at the encoder and decoder,
- Case 2. The states are known at the encoder and decoder.

The states of the wiretapper's channel are not known precisely to the transmitter or receiver. Information theoretic results on this model can be found particularly in [3], [9], [10].

This model is particularly relevant in scenarios where the channels may vary unpredictably due to environmental conditions, interference, or intentional obfuscation. The uncertainty about the wiretapper's channel introduces additional challenges in designing secure communication systems.

In dealing with compound wiretap channels, communication system designers often employ techniques from information theory, such as coding theory and channel coding, to develop strategies that maximize the secrecy capacity — the maximum achievable rate of reliable communication while keeping the information confidential from the eavesdropper under the uncertainty about the wiretapper's channel.

Designing secure communication systems for compound wiretap channels requires sophisticated analysis and optimization, taking into account the probabilistic nature of the eavesdropper's channel and balancing secrecy with the reliability and efficiency of communication.

The first information theoretic task for each channel model is to find the capacity [11]. The next task is the investigation of reliability function or E-capacity (rate-reliabilty function) suggested by E. Haroutunian [12]. In the model of wiretap channel the equivocation rate is added and the first aim is to investigate the capacity-equivocation region as well as the secrecy capacity, which was obtained in [2]. The analogy of E-capacity is the E-capacity-equivocation region $C(E, W)$, which is the closure of the set of all achievable rate-reliability-equivocation pairs $(R(E), R_e)$, where the function $R(E)$ represents the optimal dependence of the

rate R on reliability (error probability exponent) E. Some results of this investigations for various models can be found in [13], [6].

In this paper we investigate the E-capacity-equivocation region of the compound wiretap channel. The outer bound of this region is constructed in the case, when the states of the main channel are known to the sender and receiver. In [10] a similar problem was considered for the case, when the states are not known to all terminals. The practical difference in the case 2 is that the knowledge of the state affects encoding and decoding, and hence the probability of error.

2 Notations and definitions

Let us denote the parameter of the legitimate channel by $s \in \mathcal{S}$ and the parameter of the eavesdropper by $k \in \mathcal{K}$. In other words we consider the following DMC with finite input alphabet $\mathcal{X}$, finite output alphabets $\mathcal{Y}$ and $\mathcal{Z}$

$$W_{1s}^N(\mathbf{y}|\mathbf{x}) = \prod_{n=1}^{N} W_{1s}(y_n|x_n),$$

$$W_{1k}^N(\mathbf{z}|\mathbf{x}) = \prod_{n=1}^{N} W_{1k}(z_n|x_n),$$

where $x_n \in \mathcal{X}$, $y_n \in \mathcal{Y}$, $z_n \in \mathcal{Z}$, $n = 1, ..., N$, $\mathbf{x} \in \mathcal{X}^N$, $\mathbf{y} \in \mathcal{Y}^N$, $\mathbf{z} \in \mathcal{Z}^N$.

To formulate the problem, consider auxiliary random variables U and Q with values in finite sets $\mathcal{U}$ and $\mathcal{Q}$, correspondingly, that satisfy the Markov chain relationship: $Q \to U \to X \to (Y, Z)$.

Let the probability distribution (PD) of random variables (RVs) Q and U be

$$P_0 = \{P_0(q, u), q \in \mathcal{Q}, u \in \mathcal{U}\}$$

and

$$P_1 = \{P_1(x|u), x \in \mathcal{X}, u \in \mathcal{U}\}$$

be conditional PD of RV X for a given value u. Joint PD of RVs U, X we denote by

$$P_{0,1} = \{P_{0,1}(u, x) = P_0(u)P_1(x|u), u \in \mathcal{U}, x \in \mathcal{X}\}$$

and the marginal PD of X is

$$P = \{P(x) = \sum_u P_{0,1}(u, x), u \in \mathcal{U}, x \in \mathcal{X}\}.$$

We denote

$$P_1W_{1s}(y|u) = \textstyle\sum_x P_1(x|u)W_{1s}(y|x),$$

$$P_1W_{2k}(z|u) = \textstyle\sum_x P_1(x|u)W_{2k}(z|x).$$

We shall use also the PD $V = \{V(y|x), x \in \mathcal{X}, y \in \mathcal{Y}\}$.

When the state $s \in \mathcal{S}$ of the channel is **unknown** at the encoder and decoder, the N **length code** (f_N, g_N) is defined by the pair of mappings (**case 1**)

$$f_N : \mathcal{M}_N \to \mathcal{X}^N,\ g_N : \mathcal{Y}_N \to \mathcal{M}_N,$$

and if the state is **known** at the encoder and decoder (**case 2**), then

$$f_N : \mathcal{M}_N \times \mathcal{S} \to \mathcal{X}^N,\ g_N : \mathcal{Y}_N \times \mathcal{S} \to \mathcal{M}_N.$$

The **code rate** as usual is defined as

$$R(f_N, g_N) = \frac{1}{N} \log |\mathcal{M}_N|$$

(log and exp are taken to the base 2). The outer bound of the E-capacity-equivocation region in case 1 is constructed in [10]. Here we focus on the case 2.

We consider the **average error probability**, which in case 2 equals

$$e = \sup_{s \in \mathcal{S}} e(f_{N,s}, g_{N,s}, W_{1s}),$$

with

$$e(f_{N,s}, g_{N,s}, W_{1s}) = \frac{1}{|\mathcal{M}_N|} \sum_{m \in \mathcal{M}_N} W_{1s}^N \{\mathcal{Y}^N - g_{N,s}^{-1}(m) | f_{N,s}(m)\},$$

for each $s \in \mathcal{S}$. Here $g_{N,s}^{-1}(m) = \{(\mathbf{y}, s) : g_{N,s}(\mathbf{y}) = m\}$ and '–' is the operation between sets.

The secrecy level of a confidential message m at the wiretapper is measured by the **equivocation rate**, defined as

$$R_e^N = \frac{1}{N} H_{P_{01}, W_{2k}}(M | Z^N),$$

where $H_{P_{01}, W_{2k}}(M|Z^N)$ is the conditional entropy [11] with distributions P_{01}, W_{2k}. In other words, the equivocation rate indicates the eavesdropper's uncertainty about the message m given the channel outputs Z^N. Hence, the larger the equivocation rate, the higher the level of secrecy.

The rate–equivocation pair (R, R_e) is **achievable** if there exists a sequence of message sets $\mathcal{M}_N$ with $|\mathcal{M}_N| = \exp NR$ and code (f_N, g_N) such that the average error probability tends to zero as N goes to infinity, and the equivocation rate R_e satisfies

$$R_e \le \lim_{N \to \infty} \inf R_e^N.$$

The rate–equivocation pair (R, R_e) indicates the confidential rate R achieved at a certain secrecy level R_e.

The **capacity-equivocation region** $\mathcal{C}(W)$ is defined to be the closure of the set that consists of all achievable rate–equivocation pairs (R, R_e).

We investigate the E**-capacity-equivocation region** $\mathcal{C}(E, W)$, which is defined as the closure of the set that consists of all E-achievable rate–equivocation pairs $(R(E), R_e), E > 0$ with the average error probability satifying

$$e(f_N, g_N, W_{1s}) \le \exp\{-NE\}.$$

3 Formulation of results

We combine the results of compound and wiretap channels and construct the outer bound of the E-capacity-equivocation region of the compound wiretap channel for case 2. This result is formulated in the following theorem.

Theorem 1. *For $E > 0$, the outer bound for E-capacity-equivocation region of the discrete memoryless compound wiretap channel in case 2 is given by*

$$\mathcal{C}(E, W) \leq \mathcal{R}_{sp}(E, W)$$

where $\mathcal{R}_{sp}(E, W)$ equals

$$\bigcap_{s\in\mathcal{S}} \bigcup_{P_{0,1}} \left\{ \begin{array}{l} (R(E), R_e) : Q \to U \to X \to (Y, Z), \\ R(E) \leq \min\limits_{P_1V : D(P_1V||P_1W_{1s}|P_0)\leq E} I_{P_{0,1},V}(U;Y), \\ 0 \leq R_e \leq R(E), \\ R_e \leq I_{P_{0,1},W_{1s}}(U;Y|Q) - \sup\limits_{k\in\mathcal{K}} I_{P_{0,1},W_{2k}}(U;Z|Q) \end{array} \right\},$$

where $D(P_1V||P_1W_{1s}|P_0)$ denotes the divergence between conditional distributions P_1V and P_1W_1 given PD P_0 (for definition see [11]).

Proof. The proof of the first inequality is based on the method of types [14]. The second inequality is interpreted as the best state of the eavesdropper.

Corollary 1. *When $E \to 0$ the limit of this bound is the outer bound of capacity-equivocation region of the compound wiretap channel in the case 2*

$$C(W) \leq \bigcap_{s\in\mathcal{S}} \bigcup_{P_{0,1}} \left\{ \begin{array}{l} (R, R_e) : Q \to U \to X \to (Y, Z), \\ R \leq I_{P_{0,1},W_{1s}}(U;Y), \\ 0 \leq R_e \leq R, \\ R_e \leq I_{P_{0,1},W_{1s}}(U;Y|Q) - \sup\limits_{k\in\mathcal{K}} I_{P_{0,1},W_{2k}}(U;Z|Q) \end{array} \right\}.$$

Corollary 2. *In the case of degraded compound wiretap channel the outer bound will be*

$$\bigcap_{s\in\mathcal{S}} \bigcup_{P} \left\{ \begin{array}{l} (R(E), R_e) : X \to Y \to Z, \\ R(E) \leq \min\limits_{V : D(V||W_{1s}|P)\leq E} I_{P,V}(X;Y), \\ 0 \leq R_e \leq R(E), \\ R_e \leq I_{P,W_{1s}}(X;Y) - \sup\limits_{k\in\mathcal{K}} I_{P,W_{2k}}(X;Z) \end{array} \right\}.$$

The investigation will be extended by constructing the inner bounds of E-capacity-equivocation regions for cases 1 and 2.

References

1. Wyner, A. D.: The wire-tap channel. Bell System Technical Journal, **54**(8), 1355—1387 (1975).
2. Csiszár, I., Körner, J.: Broadcast channel with confidential messages. IEEE Transactions on Information Theory, **24**(3), 339—348 (1978). doi: 10.1109/TIT.1978.1055892.
3. Liang, Y., Poor, V., Shamai (Shitz), S.: Information theoretic security. Foundations and Trends in Communications and Information Theory, **5**(4-5), 355–580 (2008). DOI:10.1561/0100000036
4. Haroutunian, M.: Outer bound for E - capacity – equivocation region of the wiretap channel. In: 12th Intern. Conf. on Computer Science and Information technologies, Yerevan, Armenia (Sept. 2019) 129—131. Reprint In: IEEE Revised selected papers, 93–95 (2019). doi: 10.1109/CSITechnol.2019.8895005.
5. Haroutunian, M.: Inner bound of E - capacity – equivocation region for the generalized wiretap channel. In: 2nd CODASSCA workshop, Yerevan, Armenia 117–122 (2020).
6. Haroutunian, M.: E - capacity - equivocation region of wiretap channel. J. Universal Comput. Sci. **27**, 1222–1239 (2021). DOI:10.3897/jucs.76605
7. Wolfowitz J.: Simultaneous channels. Arch. Rational Mech. Anal. **4**, 371-386 (1960).
8. Haroutunian M.: Bounds of E - capacity for compound channels. Transactions of YSU, **3**(165), 22-29 (1987) (In Russian).
9. Liang, Y., Kramer, G., Poor, H.V. et al.: Compound wiretap channels. J Wireless Com Network, 142374 (2009). https://doi.org/10.1155/2009/142374
10. Haroutunian M.: Outer bound for E - capacity - equivocation region of compound wiretap channel. Pattern Recognition and Image Analysis, **34**(1), 137–143 (2024). DOI: 10.1134/S1054661824010073
11. Cover, T. M., Thomas, J. A.: Elements of Information Theory. 2nd edn. A Wiley-Interscience Publication, USA (2006).
12. Haroutunian, E.: E-capacity of DMC. IEEE Transactions on Information Theory, **53**(11), 4210–4220 (2007). doi: 10.1109/TIT.2007.907506.
13. 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), 97–263 (2007). doi: 10.1561/0100000008.
14. Csiszár, I.: Method of types. IEEE Transactions on Information Theory, **44**(6), 2505—2523 (1998). doi: 10.1109/18.720546.

Application of Identification Codes to the Two-Party Privacy-Preserving Record Linkage (PPRL) *

Yanling Chen[0000−0003−1603−9121]

Volkswagen Infotainment GmbH, Bochum, Germany
yanling.chen@volkswagen-infotainment.com

Abstract. In this paper, we apply the identification codes to the problem of two-party privacy-preserving record linkage (PPRL). In particular, we emphasize the advantage of our approach on the performance analysis, especially on the privacy analysis, over the classical hash-based approaches. Note for the PPRL, linkage quality is typically evaluated experimentally, whilst for privacy, there is so far no commonly accepted privacy measures available that allow an objective evaluation. Our approach of identification code provides an objective evaluation on both linkage quality and privacy based on parameters of identification codes.

Keywords: Record Linkage · Identification code · Privacy

1 Introduction

Privacy preserving record linkage (PPRL) addresses the problem of linking records that represent the same individuals across several datasets without revealing sensitive information of the individuals [4,5]. So far a variety of linkage protocols have been proposed. See a short list that includes but not limited to [6,9,11,12].

In general, proposals to PPRL can be classified into those that require a third party for performing the linkage and those that do not. The former are known as 'three-party protocols' and the latter as 'two-party protocols'. In three-party protocols, a (trusted) third party (which we call the 'linkage unit') is involved in conducting the linkage, while in two-party protocols only the two database owners participate in the PPRL process. In this paper, we put our focus on the two-party protocols.

Generally, two-party protocols start by the two database owners agreeing upon and exchanging any required information such as parameter settings, preprocessing methods, encoding or encryption methods, and any secret keys that

* The main part of the work was conducted as the author was with University of Duisburg-Essen, and was supported by the German Research Foundation under the research grant DFG 407023611.

are required, and further proceed by the secure transmission or exchange of encoded or encrypted attribute values to conduct the linkage. The final step is to derive the identified linked records.

For the privacy analysis, we assume a semi-honest threat model. More specifically, we say that a two-party PPRL protocol is *secure* in a semi-honest model when neither party is able to gain any information from the execution of the protocol, other than the information gained from the protocol's output (and the size of the other party's input).

2 Idea of Applying Identification Code to PPRL

For the standard problem of transmission, the model is shown in Fig. 1. The goal is to encode a message in a way such that after it passes through a noisy channel, the message can be successfully decoded at the receiver. It turns out that one can send messages that scale exponentially with the blocklength and have the error probability of decoding arbitrarily small (see Shannon's landmark paper [10]). For this case, error control coding provides ways of adding redundancy into messages so that the receiver can still determine the sent message correctly in spite of the noise added during the transmission.

Fig. 1. Model for standard transmission problem over channels.

In the problem of identification via channels that was introduced by Ahlswede and Dueck [1], the receiver is only interested in testing whether a particular message was sent, but the encoder does not know which message the decoder wants. The model is shown Fig. 2. In this model, the encoder sends a message a but the decoder would like to know if message b is sent. (Another interpretation in a multi-terminal setting is that, suppose that the sender's intended terminal is a; upon receiving information broadcast by the sender, each terminal b could identify whether it is the intended recipient.) It turns out that one can design systems such that the number of different messages/terminals one can identify grows doubly exponentially with the blocklength. For this case, errors are considered in terms of false identification and missed identification; and the idea behind the optimal coding strategy is to map each message into a list of codewords and the encoder selects one randomly; as long as the fraction of the pairwise overlap of these lists is small, the error probabilities will be small.

Immediately we notice the similarity between the problem of identification via channels and the problem of record linkage, especially if we consider the scenario of two-party PPRL with one data holder A as the encoder, the other data holder B as the decoder, and each record corresponding to a message/terminal. See Fig. 3 for the linkage model. For the record comparison, data holder A sends

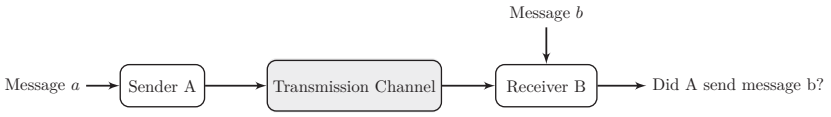


Fig. 2. Model for identification problem over channels.

information about record r_a (e.g., an anonymized form of record r_a) to data holder B. Data holder B tries to identify whether it is a match with record r_b he has (e.g., via comparison between the anonymized form of r_a with an similarly anonymized record r_b). Clearly, once two data holders agree on using an identification code for the two-party PPRL, the encoding procedure is conducted at one data holder to anonymize its records; whilst the decoding procedure is employed at the other data holder to link the record pairs.

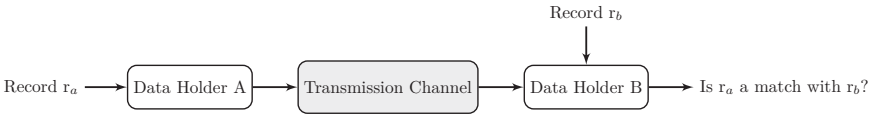


Fig. 3. Model for record linkage problem in a 2-party PPRL protocol.

3 Preliminaries

Before proceeding, we provide some definitions that will be useful in the problem setup, especially to distinguish between the definition of codes used for error control and codes for identification. In general, we assume a channel with input alphabet $\mathcal{A}$ and output alphabet $\mathcal{B}$. We also notice that the scenario of PPRL corresponds to the noiseless channel in which $\mathcal{A} = \mathcal{B}$.

3.1 Transmission code

For the transmission problem, we use (n, L, λ) to denote a transmission code which satisfies

$$\Pr\{s \text{ is decoded at receiver} \mid s \text{ is sent by transmitter}\} \geq 1 - \lambda,$$

for each message s, where each codeword has length n and there are L messages. The *rate* of an (n, L, λ) transmission code is $\frac{1}{n} \log L$.

3.2 Identification plus Transmission code (IT code)

Han and Verdú [7] also introduced the model of identification plus transmission code (IT code), where a central station wishes to transmit one of the M messages

to one of the N terminals (suppose that codeword $f(a,m)$, of length n, is sent for message m to terminal a). Upon receiving the codeword, each terminal decides whether it is the intended recipient of the message and if so it decodes the message. The decoding reliability of which is measured by (λ_1, λ_2) as follows:

1. for each terminal a :

$$\Pr\left\{\begin{array}{l|l} a \text{ decides that it is the intended} & a \text{ is the intended recipient,} \\ \text{recipient} & m \text{ is transmitted} \end{array}\right\} \geq 1-\lambda_1,$$

2. for any pair of terminals $b \neq a$:

$$\Pr\left\{\begin{array}{l|l} b \text{ decides that it is the intended} & a \text{ is the intended recipient,} \\ \text{recipient} & m \text{ is transmitted} \end{array}\right\} \leq \lambda_2,$$

where the probability is taken over all codewords for terminal a in both equations. The *rate-pair* of an $(n, N, M, \lambda_1, \lambda_2)$ IT code is $(\frac{1}{n}\log M, \frac{1}{n}\log\log N)$.

3.3 Identification code (ID code)

Given any $(n, N, M, \lambda_1, \lambda_2)$ IT code, one can immediately construct an $(n, N, \lambda_1, \lambda_2)$ ID code by choosing m randomly over $\{1, \cdots, M\}$ in the encoding function $f(a,m)$ for each $a = 1, \cdots, N$. We use $(n, N, \lambda_1, \lambda_2)$ to denote the obtained identification code (ID code), which satisfies

1. for each terminal a :

$$\Pr\{a \text{ decides that it is intended} \mid a \text{ is the intended recipient}\} \geq 1-\lambda_1,$$

2. for any pair of terminals $b \neq a$:

$$\Pr\{b \text{ decides that it is intended} \mid a \text{ is the intended recipient}\} \leq \lambda_2,$$

where the probability is taken over all codewords for terminal a in both equations, each codeword has length n and there are N terminals. The *rate* of an $(n, N, \lambda_1, \lambda_2)$ ID code is $\frac{1}{n}\log\log N$.

4 Relationship between λ_1, λ_2 and precision, sensitivity/recall, specificity

For each encoding and decoding that corresponds to a record comparison in two-party PPRL, the decoding result can be assigned into the following 4 categories: True positives (TP), False positives (FP), True negatives (TN), False negatives (FN). In particular, we have

$$\begin{aligned}
\Pr\{\text{TP}\} &= \Pr\{b \text{ decides it is intended } \&\ b = a \mid a \text{ is the intended recipient}\};\\
\Pr\{\text{FP}\} &= \Pr\{b \text{ decides it is intended } \&\ b \neq a \mid a \text{ is the intended recipient}\};\\
\Pr\{\text{TN}\} &= \Pr\{b \text{ decides it is not intended } \&\ b \neq a \mid a \text{ is the intended recipient}\};\\
\Pr\{\text{FN}\} &= \Pr\{b \text{ decides it is not intended } \&\ b = a \mid a \text{ is the intended recipient}\}.
\end{aligned}$$

Note that if an $(n, N, \lambda_1, \lambda_2)$ ID code is employed, we have

$$\begin{aligned}&\Pr\{b \text{ decides it is intended} \mid a \text{ is the intended recipient } \&\ b = a\ \} \geq 1 - \lambda_1;\\ &\Pr\{b \text{ decides it is intended} \mid a \text{ is the intended recipient } \&\ b \neq a\ \} \leq \lambda_2.\end{aligned}$$

Furthermore, we notice that

$$\begin{aligned}&\Pr\{b \text{ decides that it is intended} \mid a \text{ is the intended recipient } \&\ b = a\ \}\\ &\quad = \frac{\Pr\{b \text{ decides that it is intended } \&\ b = a \mid a \text{ is the intended recipient}\}}{\Pr\{b = a \mid a \text{ is the intended recipient}\}}\\ &\quad = \frac{\Pr\{\mathrm{TP}\}}{\Pr\{\mathrm{TP}\} + \Pr\{\mathrm{FN}\}}\\ &\quad \geq 1 - \lambda_1;\\ &\Pr\{b \text{ decides that it is intended} \mid a \text{ is the intended recipient } \&\ b \neq a\ \}\\ &\quad = \frac{\Pr\{b \text{ decides that it is intended } \&\ b \neq a \mid a \text{ is the intended recipient}\}}{\Pr\{b \neq a \mid a \text{ is the intended recipient}\}}\\ &\quad = \frac{\Pr\{\mathrm{FP}\}}{\Pr\{\mathrm{FP}\} + \Pr\{\mathrm{TN}\}} = 1 - \frac{\Pr\{\mathrm{TN}\}}{\Pr\{\mathrm{FP}\} + \Pr\{\mathrm{TN}\}}\\ &\quad \leq \lambda_2.\end{aligned}$$

Recall that in the context of record linkage,

$$\begin{aligned}\text{Precision} &= \frac{\Pr\{\mathrm{TP}\}}{\Pr\{\mathrm{TP}\} + \Pr\{\mathrm{FP}\}};\\ \text{Sensitivity/Recall} &= \frac{\Pr\{\mathrm{TP}\}}{\Pr\{\mathrm{TP}\} + \Pr\{\mathrm{FN}\}}.\\ \text{Specificity} &= \frac{\Pr\{\mathrm{TN}\}}{\Pr\{\mathrm{TN}\} + \Pr\{\mathrm{FP}\}}.\end{aligned}$$

Then for the employed $(n, N, \lambda_1, \lambda_2)$ ID code, we obtain

$$\text{Precision} \geq \frac{\Pr\{\mathrm{TP}\}}{\Pr\{\mathrm{TP}\} + \lambda_2}; \quad \text{Sensitivity/Recall} \geq 1 - \lambda_1; \quad \text{Specificity} \geq 1 - \lambda_2.$$

In general, smaller λ_1, λ_2 imply better precision, sensitivity/recall and specificity scores.

5 Construction of ID codes

5.1 Binary constant-weight code (BCWC)

Definition 5.1. *An (L, N, M, K) binary constant-weight code is a set of N binary strings of length L and Hamming weight M such that the pairwise overlap (maximum number of coincident 1's between any two codewords) does not exceed K.*

Any (L, N, M, K) binary constant-weight code can be described by an $N \times M$ incidence matrix on $\{1, \cdots, L\}$ s.t. the row $(s(a,1), \cdots, s(a,M))$ gives the locations of the M 1's in the ath codeword, for every $a \in \{1, \cdots, N\}$. Define

$$\beta = \frac{\log M}{\log L}, \quad \rho = \frac{\log \log N}{\log L}, \quad \mu = \frac{K}{M},$$

where β is called the *weight factor*, ρ is called the *second-order rate* (as opposed to the first-order rate $\frac{1}{L}\log N$), and μ is called the *overlap fraction* of the binary constant-weight code.

$$\begin{array}{c} \text{Code} \\ \begin{bmatrix} 1 \\ \vdots \\ a \\ \vdots \\ N \end{bmatrix} \end{array} \quad \begin{array}{c} 1 \ 2 \ \cdots \ \cdots \cdot \ L \\ \begin{bmatrix} 1 & 0 & \cdots & 0 \cdots & 0 & 1 \\ \vdots & \vdots & \cdots & \vdots \cdots & \vdots & \vdots \\ 0 & 0 & \cdots & 0 \cdots & 1 & 1 \\ \vdots & \vdots & \cdots & \vdots \cdots & \vdots & \vdots \\ 1 & 1 & \cdots & 0 \cdots & 0 & 0 \end{bmatrix} \\ \text{Each row has } M \text{ 1's,} \\ \text{Between rows} \le K \text{ overlapping 1's.} \end{array} \quad \Leftrightarrow \quad \begin{array}{c} \text{Matrix} \\ \begin{bmatrix} 1 \\ \vdots \\ a \\ \vdots \\ N \end{bmatrix} \end{array} \quad \begin{array}{c} 1 \cdots \ m \ \cdots M \\ \begin{bmatrix} \cdots & \cdot & \cdots \\ \vdots \cdots & \vdots & \cdots \vdots \\ \cdots & s(a,m) & \cdots \\ \vdots \cdots & \vdots & \cdots \vdots \\ \cdots & \cdot & \cdots \end{bmatrix} \\ \text{where } a \in \{1, \cdots, N\}, \\ m \in \{1, \cdots, M\}, \\ s(a,m) \in \{1, \cdots, L\}. \end{array}$$

5.2 Construction of ID codes via BCWC

Verdú and Wei [13, Proposition 1] showed that an IT code can be obtained by concatenating a transmission code with a binary constant-weight code.

More specifically, we consider the special case where the underlying channel is noiseless, that is, $\mathcal{A} = \mathcal{B}$ and no errors would occur during the transmission from the sender to the receiver (thus $\lambda = 0$). For such a case, a one-to-one mapping can be used for the encoding and decoding purpose. For instance, taking $\mathcal{A} = \mathcal{B} = \{0, 1\}$ and $n = \lfloor \log_2 L \rfloor$, one can define for this transmission code an encoding function $\phi(\cdot)$ that maps s to the binary representation of $s - 1$ for $s \in \{1, \cdots, L\}$ in n bits. Since the channel is noiseless, the decoding function could be simply the inverse mapping $\phi^{-1}(\cdot)$.

Algorithm 1 Encoding of an $(n, N, 0, \mu)$ ID code from the $(n, N, M, 0, \mu)$ IT code via an $(n, L, 0)$ transmission code with an encoding function $\phi(\cdot)$ & an $(L, N, M, \mu M)$ binary constant-weight code with an incidence matrix S.

Input: Intended terminal a. ▷ $a \in \{1, \cdots, N\}$.

1: Randomly choose m over $\{1, \cdots, M\}$.

2: Compute codeword $c := \phi(S(a, m))$. ▷ $\phi(\cdot)$ is the encoding function of the $(n, L, 0)$ transmission code.

▷ $S(a, m)$ is the element of S in the a-th row, m-th column.

Output: Codeword c.

According to [13, Proposition 1], an $(n, N, M, 0, \mu)$ IT code can be obtained by concatenating an $(n, L, 0)$ transmission code with an $(L, N, M, \mu M)$ binary constant-weight code. Furthermore, an $(n, N, 0, \mu)$ ID code could be obtained from the $(n, N, M, 0, \mu)$ IT code, the encoding and decoding of which are described in Algorithm 1 and 2, respectively.

Algorithm 2 Decoding of an $(n, N, 0, \mu)$ ID code from the $(n, N, M, 0, \mu)$ IT code via an $(n, L, 0)$ transmission code with a decoding function $\phi^{-1}(\cdot)$ & an $(L, N, M, \mu M)$ binary constant-weight code with an incidence matrix S.

Input: Terminal b and codeword c'. ▷ $b \in \{1, \cdots, N\}$.
1: Compute $s' := \phi^{-1}(c')$. ▷ $\phi^{-1}(\cdot)$ is the decoding function of the $(n, L, 0)$ transmission code.
2: **if** $s' \in S(b, \star)$ **then** ▷ $S(b, \star)$ is the set of elements in the b-th row of S.
3: flag ← TRUE; ▷ b declares that it is the intended recipient.
4: **else**
5: flag ← FALSE; ▷ b declares that it is not the intended recipient.
6: **end if**
Output: flag.

5.3 Construction of BCWC via Universal Hash Functions

Let $\mathcal{X}$ and $\mathcal{Y}$ be finite sets such that $|\mathcal{X}| \geq |\mathcal{Y}|$, and $\mathcal{H}$ be a set of functions such that $h : \mathcal{X} \to \mathcal{Y}$ for each $h \in \mathcal{H}$.

Definition 5.2. *We say that $\mathcal{H}$ is an ϵ-almost strongly universal (ϵ-ASU) class of hash functions provided that the following two conditions are satisfied:*

- *for any $x \in \mathcal{X}, y \in \mathcal{Y}$, there exist exactly $\frac{|\mathcal{H}|}{|\mathcal{Y}|}$ functions $h \in \mathcal{H}$ such that $h(x) = y$;*
- *from any two distinct elements $x_1, x_2 \in \mathcal{X}$ and for any two (not necessarily distinct) elements $y_1, y_2 \in \mathcal{Y}$, there exist at most $\epsilon\frac{|\mathcal{H}|}{|\mathcal{Y}|}$ functions $h \in \mathcal{H}$ such that $h(x_i) = y_i, i = 1, 2$.*

Let $\mathcal{H}$ be an ϵ-ASU class of hash functions from $\mathcal{X}$ to $\mathcal{Y}$. The incidence matrix of $\mathcal{H}$ is an $|\mathcal{X}||\mathcal{Y}| \times |\mathcal{H}|$ binary matrix defined by

$$((x, y), h)\text{-th element} = \mathbb{1}_{\{h(x)=y\}} = \begin{cases} 1, \text{ if } h(x) = y; \\ 0, \text{ otherwise.} \end{cases}$$

Then the incidence matrix of $\mathcal{H}$ is an (L, N, M, K) binary constant-weight code with

$$L = |\mathcal{H}|, \quad N = |\mathcal{X}||\mathcal{Y}|, \quad M = \frac{|\mathcal{H}|}{|\mathcal{Y}|}, \quad K = \epsilon\frac{|\mathcal{H}|}{|\mathcal{Y}|},$$

and the overlap factor $u = \frac{K}{M} = \epsilon$.

$$\begin{array}{c|c|c} \text{Code via } \mathcal{H} & h_1 \cdots \quad h_i \quad \cdots h_{|\mathcal{H}|} & \\ \hline \begin{array}{c}1\\ \vdots \\ (x,y)\\ \vdots \\ |\mathcal{X}||\mathcal{Y}|\end{array} & \begin{array}{ccc}\cdots & \cdot & \cdots \\ \vdots & \vdots & \vdots \\ \cdots & \mathbb{1}_{\{h_i(x)=y\}} & \cdots \\ \vdots & \vdots & \vdots \\ \cdots & \cdot & \cdots\end{array} & \begin{array}{r}\text{where}\\ x \in \mathcal{X},\\ y \in \mathcal{Y},\\ \mathcal{H} = \{h_i | i \in \{1, \cdots, |\mathcal{H}|\}\}.\end{array} \end{array}$$

5.4 Construction of ϵ-ASU class of hash functions

Let q be a prime power and let $1 \leq k \leq q$. Let $\mathcal{X} = \{(a_1, \cdots, a_k) | a_i \in \mathbb{GF}(q)\}$ and $\mathcal{Y} = \{\text{the elements of } \mathbb{GF}(q)\}$.

den Boer [3] described the following ϵ-ASU class of hash functions from $\mathcal{X}$ to $\mathcal{Y}$:

Definition 5.3. *For* $\forall e_0, e_1 \in \mathbb{GF}(q)$, *let*

$$h_{(e_0,e_1)}(a_1, \cdots, a_k) = e_0 + a_1 e_1 + \cdots + a_k e_1^k.$$

Then $\mathcal{G}(q,k) = \{h_{(e_0,e_1)}\}$ *is an* ϵ*-ASU class of hash functions from* $\mathcal{X}$ *to* $\mathcal{Y}$ *such that* $|\mathcal{G}(q,k)| = q^2$ *and* $\epsilon = \frac{k}{q}$.

Such an ϵ-ASU class of hash functions implies an $(L, N, M, \mu M)$ binary constant-weight code with

$$L = |\mathcal{G}(q,k)| = q^2, \quad N = |\mathcal{X}||\mathcal{Y}| = q^{k+1}, \quad M = \frac{|\mathcal{G}(q,k)|}{|\mathcal{Y}|} = q, \quad u = \epsilon = \frac{k}{q}.$$

Moreover, if we consider an $(n, L, 0)$ transmission code with $n = 2$ and $L = q^2$ (i.e., each codeword is of length 2 over $\mathbb{GF}(q)$); and an $(L, N, M, \mu M)$ binary constant-weight code constructed by $\mathcal{G}(q,k)$ as defined above with $L = q^2, N = q^{k+1}, M = q$ and $\mu = \frac{k}{q}$, then we obtain an $(2, q^{k+1}, 0, \frac{k}{q})$ ID code according to [13, Proposition 1]. Note that this code could identify $N = q^{k+1}$ terminals. Each terminal can be indexed by $(\mathbf{a}, \alpha)$, where $\mathbf{a} = (a_1, \cdots, a_k)$ with $a_i \in \mathbb{GF}(q)$ for $i = 1, \cdots, k$ and α is an element of $\mathbb{GF}(q)$.

$$\begin{array}{c|c|c} \text{Code via } \mathcal{G}(q,k) & h_1 \cdots h_{(e_0,e_1)} \cdots h_{|\mathcal{G}(q,k)|} & \\ \hline \begin{array}{c}1\\ \vdots \\ (\mathbf{a},\alpha)\\ \vdots \\ |\mathcal{X}||\mathcal{Y}|\end{array} & \begin{array}{ccc}\cdots & \cdot & \cdots \\ \vdots & \vdots & \vdots \\ \cdots & \mathbb{1}_{\{h_{(e_0,e_1)}(\mathbf{a})=\alpha\}} & \cdots \\ \vdots & \vdots & \vdots \\ \cdots & \cdot & \cdots\end{array} & \begin{array}{r}\text{where}\\ \mathbf{a} = (a_1, \cdots, a_k) \in \mathcal{X} = \mathbb{GF}(q)^k,\\ \alpha \in \mathcal{Y} = \mathbb{GF}(q),\\ \mathcal{G}(q,k) = \{h_{(e_0,e_1)} | e_0, e_1 \in \mathbb{GF}(q)\}.\end{array} \end{array}$$

Algorithm 3 Encoding of the $(2, q^{k+1}, 0, \frac{k}{q})$ ID code

Input: Intended terminal $(\mathbf{a}, \alpha)$. ▷ $\mathbf{a} = (a_1, \cdots, a_k) \in \mathbb{GF}(q)^k$ and $\alpha \in \mathbb{GF}(q)$.
1: Randomly choose e_1 over $\mathbb{GF}(q)$. ▷ A random choice from those q choices of (e_0, e_1) s.t. $h_{(e_0,e_1)}(\mathbf{a}) = \alpha$.
2: Compute e_0 s.t. $e_0 + a_1 e_1 + \cdots + a_k e_1^k = \alpha$.
Output: Codeword (e_0, e_1). ▷ $e_0, e_1 \in \mathbb{GF}(q)$.

The encoding and decoding of this specific ID code instance are described in Algorithm 3 and 4, respectively. Comparing with the general ID codes constructed by using binary constant-weight codes, this specific construction offers a few advantages. It provides efficient encoding and decoding algorithms (i.e., reduced computational complexity for each encoding-decoding procedure). Besides, there is also no need to store the N by M incidence matrix S at both the encoder and decoder (thus saving the storage cost at both sides).

Algorithm 4 Decoding of the $(2, q^{k+1}, 0, \frac{k}{q})$ ID code

Input: Terminal $(\mathbf{b}, \beta)$ and codeword (e_0, e_1). ▷ $\mathbf{b} = (b_1, \cdots, b_k) \in \mathbb{GF}(q)^k$ and $\beta \in \mathbb{GF}(q)$.
1: **if** $e_0 + b_1 e_1 + \cdots + b_k e_1^k == \beta$ **then** ▷ Check whether $h_{(e_0,e_1)}(\mathbf{b}) = \beta$ hold or not.
2: flag ← TRUE; ▷ $(\mathbf{b}, \beta)$ declares that it is the intended recipient.
3: **else**
4: flag ← FALSE; ▷ $(\mathbf{b}, \beta)$ declares that it is not the intended recipient.
5: **end if**
Output: flag.

6 Using ID codes in two-party PPRL

For a two-party PPRL, suppose that data holder A sends the anonymized data set to data holder B; and data holder B conducts the linkage based on the anonymized data set from A and its own data set. Then an ID code could be used in a two-party PPRL protocol by considering each record as a terminal. In more detail, the record anonymization procedure at the data holder A for each record is corresponding to the encoding procedure by taking the record as input; whilst the record linkage procedure at the data holder B for each record pair is corresponding to the decoding procedure by taking received codeword and record at B as input. The decoding outputs a TRUE or FALSE to flag a match or non-match of the record pair.

More specifically, an $(n, N, 0, \mu)$ ID code can be used to link any two records from N different entities. And, for each record, a codeword of length n needs to be transmitted from data holder A to data holder B. According to the discussion in

Sec. 4, a simple application of the code (for exact matching) offers a performance for the two-party PPRL with Sensitivity/Recall = 1 and a Specificity $\geq 1-\mu$.

6.1 Applying ID code to two-party PPRL

Suppose that each person has a unique identification number across the data sets to be linked. Then the linkage can be conducted based on exact matching. That is, two records which share the same identification number will be considered as a match.

Assume that an $(n, N, 0, \mu)$ ID code (and other cryptographic parameters, if any) is known to both data holder A and data holder B. Suppose this ID code is obtained by concatenating an $(n, L, 0)$ transmission code with an $(L, N, M, \mu M)$ binary constant-weight code (that is described by the N by M incidence matrix S). Then the two-party protocol based on such an ID code is described as follows.

At data holder A, for each record i,

- the identification number is first transformed into a private match-key, say mk_i;
  To apply the ID $(n, N, 0, \mu)$ code, $mk_i \in \{1, \cdots, N\}$ is prepared by taking an appropriate transformation.
- taking mk_i as the input to the encoding procedure of the agreed $(n, N, 0, \mu)$ ID code as described in Algorithm 1, the output gives an anonymized form of the record i.
  Note that for the PPRL scenario, one can simply use the identity function as the encoding and decoding function of the transmission code (since a noiseless transmission channel is assumed). Then the output codeword for record i is $S(mk_i, r_i)$, where $S(mk_i, r_i)$ is the element of matrix S in the mk_i-th row and r_i-th column, taking value over $\{1, \cdots, L\}$; and r_i is randomly chosen in the encoding procedure, taking value over $\{1, \cdots, M\}$.

After this is done for every record, data holder A sends the list of $\{S(mk_i, r_i)\}$ to data holder B. Since $S(mk_i, r_i)$ takes values over $\{1, \cdots, L\}$, for each record $\lfloor \log_2 L \rfloor$ bits need to be transmitted to data holder B for linkage.

At the data holder B, the matched records could be identified due to possessing of the same private match-key mk'. In particular, data holder B employs the decoding procedure (as described in Algorithm 2) of the agreed $(n, N, 0, \mu)$ ID code. If $mk' = mk_i$, then it is clear that $S(mk_i, r_i) \in S(mk', \star)$, where $S(mk', \star)$ is the set of elements in the mk'-th row of S. The decoding algorithm will return flag = TRUE and this results in a match.

In general, this scheme has a Sensitivity/Recall = 1 and Specificity $\geq (1-\mu)$.

7 Performance analysis for 2-party PPRL with single match-key

Assume data holder A and data holder B each hold a set of records, independently chosen from the population. To conduct the record linkage, data holder

A and data holder B agree on using an $(n, N, 0, \mu)$ ID code, which is obtained by concatenating an $(n, L, 0)$ transmission code with an $(L, N, M, \mu M)$ binary constant-weight code (that is described by the N by M incidence matrix S). Straightforwardly, we have the following interpretation of the parameters (if applying this code to 2-party PPRL with single match-key):

- N : the maximal number of entities that can be identified, i.e., *identification capacity.*
- n : the length of the codeword, i.e., the number of digits that data holder A needs to transmit to data holder B for each record, i.e., *transmission cost.*
- M : the maximal number of anonymized forms that can be generated for each record, i.e., *randomness for anonymization.*
- μ : probability of false identification that leads to Precision $\geq \frac{\Pr\{\mathrm{TP}\}}{\Pr\{\mathrm{TP}\}+\mu}$ and Specificity $\geq 1-\mu$.

Let $\mathcal{C}$ denote the $(n, N, 0, \mu)$ ID code shared at both data holders. For each record r_A, where $\mathrm{r}_A \in \{1, \cdots, N\}$, data holder A sends a codeword $c(\mathrm{r}_A)$, which is an anonymized form of r_A, to data holder B. A natural question is that, how much information data holder B gains on record r_A by receiving $c(\mathrm{r}_A)$? This *information gain* at data holder B is measured by

$$H(\mathrm{r}_A|\mathrm{r}_B, \mathcal{C}) - H(\mathrm{r}_A|c(\mathrm{r}_A), \mathrm{r}_B, \mathcal{C}),$$

where $H(\cdot)$ is the entropy function. In other words, this is the *information leakage* about r_A from data holder A. The less it is, the better is the privacy. Besides, a normalized definition is the *relative information gain* that is measured by

$$\frac{H(\mathrm{r}_A|\mathrm{r}_B, \mathcal{C}) - H(\mathrm{r}_A|c(\mathrm{r}_A), \mathrm{r}_B, \mathcal{C})}{H(\mathrm{r}_A|\mathcal{C})}.$$

Proposition 7.1. *The information gain at data holder B is* $\leq \log_2 \frac{L}{M}$ *bits.*

Proof. First we note that

$$\begin{aligned} H(\mathrm{r}_A|c(\mathrm{r}_A), \mathrm{r}_B, \mathcal{C}) &= H(\mathrm{r}_A, c(\mathrm{r}_A)|\mathrm{r}_B, \mathcal{C}) - H(c(\mathrm{r}_A)|\mathrm{r}_B, \mathcal{C}) \\ &= H(\mathrm{r}_A|\mathrm{r}_B, \mathcal{C}) + H(c(\mathrm{r}_A)|\mathrm{r}_A, \mathrm{r}_B, \mathcal{C}) - H(c(\mathrm{r}_A)|\mathrm{r}_B, \mathcal{C}). \end{aligned}$$

Then the information gain at data holder B is

$$\begin{aligned} H(\mathrm{r}_A|\mathrm{r}_B, \mathcal{C}) - H(\mathrm{r}_A|c(\mathrm{r}_A), \mathrm{r}_B, \mathcal{C}) &= H(c(\mathrm{r}_A)|\mathrm{r}_B, \mathcal{C}) - H(c(\mathrm{r}_A)|\mathrm{r}_A, \mathrm{r}_B, \mathcal{C}) \\ &= H(c(\mathrm{r}_A)|\mathrm{r}_B, \mathcal{C}) - H(c(\mathrm{r}_A)|\mathrm{r}_A, \mathcal{C}) \\ &\leq \log_2 L - \log_2 M = \log_2 \frac{L}{M}, \end{aligned}$$

where the last inequality is due to the facts that

1. $H(c(\mathrm{r}_A)|\mathrm{r}_A, \mathcal{C}) = \log_2 M$. This is due to the fact that given the code $\mathcal{C}$ (with incidence matrix S), $c(\mathrm{r}_A)$ is chosen randomly from the locations of M 1's in the row of matrix S that is corresponding to r_A;

2. $H(c(\mathrm{r}_A)|\mathrm{r}_B, \mathcal{C}) \leq H(c(\mathrm{r}_A)) \leq \log_2 L$, since conditioning reduces entropy and $c(\mathrm{r}_A)$ takes value over $\{1, \cdots, L\}$.

So the information gain at data holder B is $\leq \log_2 \frac{L}{M}$ bits.

So far we have discussed the performance if we employ an $(n, N, 0, \mu)$ ID code, which is obtained by concatenating an $(n, L, 0)$ transmission code with an $(L, N, M, \mu M)$ binary constant-weight code. Moreover, concrete code realizations may lead to different costs on the code sharing, computations at anonymization and linkage, and privacy performance.

7.1 Shared parameters

To apply an ID code to a two-party PPRL protocol, the ID code needs to be shared by data holder A and data holder B for the purpose of the successful linkage. For the ID code is constructed via an $(L, N, M, \mu M)$ binary constant-weight code that is described by an N by M incidence matrix S (as discussed in Sec. 5), the storage cost for the incidence matrix S could be expensive, especially if both data holders have to store the codebook and when N is large (which is supposed to be at least as large as the size of the merged data set A and B).

Some specific constructions of the $(L, N, M, \mu M)$ binary constant-weight codes could avoid such problems. For instance, the construction via ϵ-ASU class of universal hash functions as discussed in Sec. 5.4 is an attractive option, with which there is no need to store the incidence matrix S at both data holders to facilitate the encoding and decoding procedures (see Algorithms 3 and 4, S is not needed for both the encoding and decoding procedures).

7.2 Computational cost

For the ID code constructed via an $(L, N, M, \mu M)$ binary constant-weight code that is described by an N by M incidence matrix S (as discussed in Sec. 5), the computation cost for the anonymization or linkage could be expensive, especially when M is large, according to Algorithms 3 and 4.

Again, some specific constructions of the $(L, N, M, \mu M)$ binary constant-weight codes could offer efficient encoding and decoding of the ID code (and thus efficient anonymization and linkage). Here the construction via ϵ-ASU class of universal hash functions as discussed in Sec. 5.4 is again an attractive option also in this aspect. Especially for the decoding procedure, instead of checking whether the received codeword belongs to a set of M elements, only one equality needs to be checked (see Algorithm 2 and 4).

7.3 Some concrete choices of ID codes

Consider the following two concrete ID code families:

1. $\mathcal{C}_1$: the $(2, q^{k+1}, 0, \frac{k}{q})$ ID code as we discussed in Sec. 5.4;

2. $\mathcal{C}_2$: the $(k+2, q^{kq^t+1}, 0, \mu)$ ID code with $\mu = \frac{k}{q} + \frac{1}{q^{k-t}} - \frac{1}{q^k}$ and $t < k$ as proposed in [8].

Some observations can be made as shown in Table 1.

Table 1. Performance comparison between $\mathcal{C}_1$ and $\mathcal{C}_2$ in 2-party PPRL with unique match-key

Performance	$(2, q^{k+1}, 0, \frac{k}{q})$ **ID code**	$(k+2, q^{kq^t+1}, 0, \frac{k}{q} + \frac{1}{q^{k-t}} - \frac{1}{q^k})$ **ID code**
Identification capacity	q^{k+1}	q^{kq^t+1}
Transmission cost	$2\log_2 q$ bits	$(k+2)\log_2 q$ bits
Randomness for anonymization	$\log_2 q$ bits	$(k+1)\log_2 q$ bits
Prob. missed identification	0	0
Prob. false identification	$\leq \frac{k}{q}$	$\leq \frac{k}{q} + \frac{1}{q^{k-t}} - \frac{1}{q^k}$
Information gain	$\leq \log_2 q$ bits	$\leq \log_2 q$ bits

If the records are uniformly distributed over $\{1, \cdots, N\}$, then code $\mathcal{C}_1$ offers a relative information gain $\leq \frac{1}{k+1}$; whilst code $\mathcal{C}_2$ offers a relative information gain $\leq \frac{1}{kq^t+1}$.

It is worth mentioning that the probability of false identification μ is the probability to link a random pair of records erroneously. Suppose that the data holder A has a dataset of n_A records, whilst data holder B has a dataset of n_B records, and there are n_M true matches between these two datasets. Then an estimate for the number of erroneously linked pairs of records is given by $u \cdot (n_A - n_M) \cdot (n_B - n_M)$ and upper bounded by $u \cdot n_A n_B$. For instance, if ID code $\mathcal{C}_1$ is used for the record linkage, in order to obtain a low homonym error rate, (k, q) can be chosen such that $\frac{k}{q} \cdot n_A n_B \leq 1$, which leads to $q \geq k n_A n_B$.

7.4 Comparison to hash based two-party PPRL

7.4.1 Data We are using the 'voter ID data' that is available at `https://dl.ncsbe.gov/index.html?prefix=data/`. The data sets at data holder A and B are generated by independently samplying the 'voter ID data' with sampling size ranging from 500 to 2500 records.

7.4.2 Linkage For comparison purpose, the first and last name attributes are merged into a single string per record. The following 3 methods are considered:

- *plain text*, i.e., data holder A sends the strings in plain text to data holder B;
- *Hash*, i.e., data holder A first hashes the strings and then sends them to data holder B;
- *ID code*, i.e., data holder A applies the ID code to (the hash of) each string and then sends the coded version to data holder B.

No blocking are employed. The aim to simply compare the anonymization time by using Hash and ID code; and linkage time by using these three different methods.

7.4.3 Implementation details All 3 different encoding methods and linkage are handled using R 4.0.1. In particular, the R package fastdigest [2] is used to create 128-bit hashes of randomly drawn strings. For the ID code, we choose different (k, q) values in code $\mathcal{C}_1$ to illustrate how they impact on the performance.

7.4.4 Linkage quality measures For simplicity, we take the plain text comparison as the golden standard and only consider the exact matching (with a single match-key).

Setting different parameters (k, q), one can obtain different ID code. As one can see from Fig. 4, q plays an important role in the linkage quality, which is reflected by MPR (mean of precision and recall). In general, the larger is q, the better is the MPR. Especially, the largest choice $q = 82589933$ is larger than kn_An_B for the choices of k, n_A, n_B in the experiments. As expected, it gives the best performance on the accuracy.

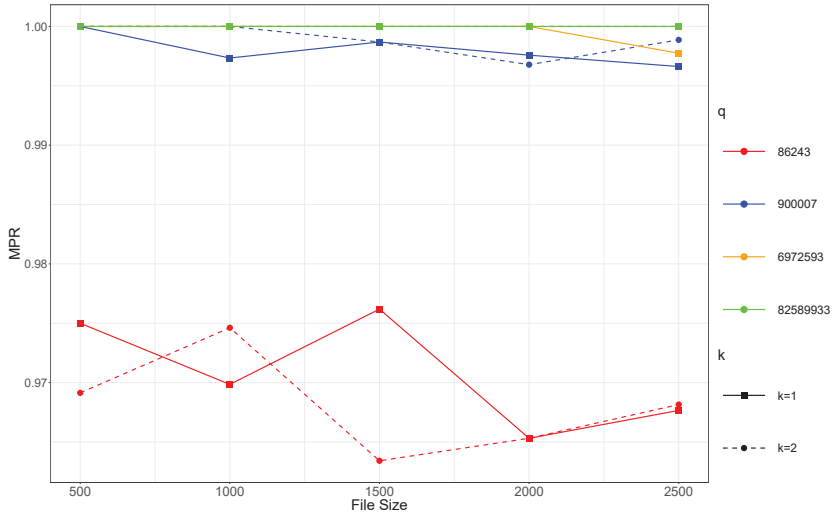


Fig. 4. MPR for ID code with different (q, k) settings.

7.4.5 Complexity In general, it takes data holder A (whose task is mainly the anonymization, which has a linear complexity) much shorter time than data holder B (whose task is mainly the linkage, which has a square complexity) in a two-party PPRL protocol. For each data holder has a data set with around 1000 records, anonymization takes about 10^{-4} minutes; while linkage takes about 10^{-2} minutes. Now let us consider only the hash and ID code. Both of methods aim to provide a certain degree of privacy. As one can observe from Fig. 6, for both the anonymization and the linkage phase, using hashing is more efficient than ID code.

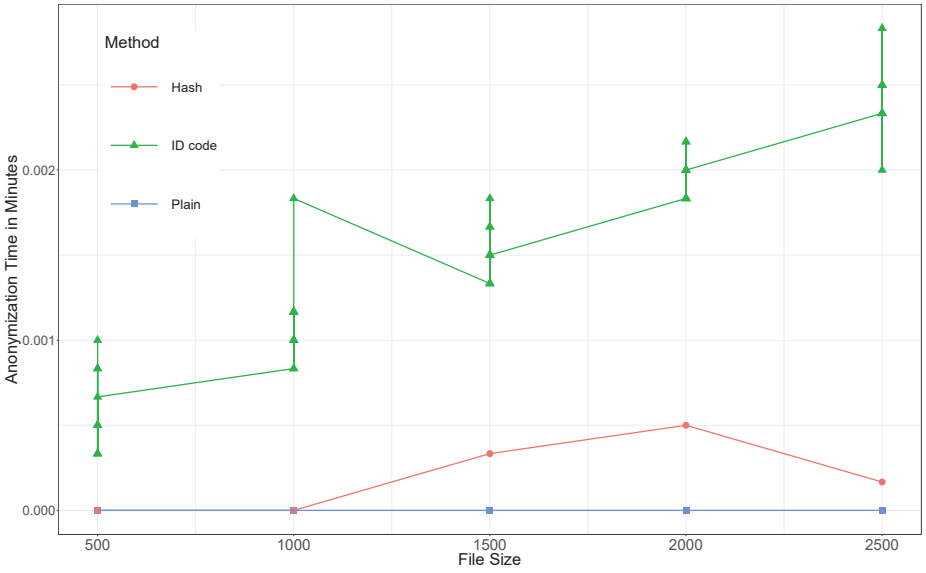


Fig. 5. Model for standard transmission problem over channels.

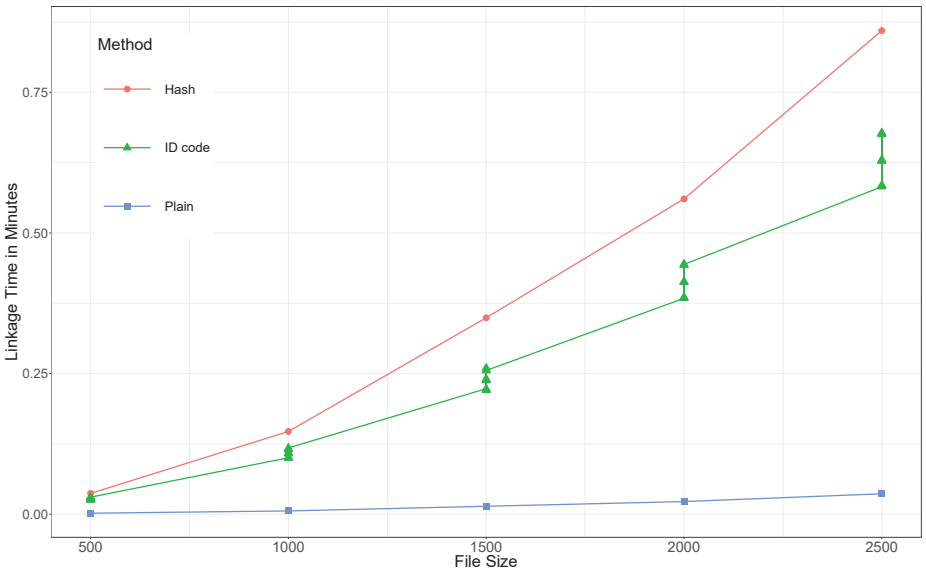


Fig. 6. Model for standard transmission problem over channels.

8 Conclusion

In this paper, we show that the identification problem over channels in Information Theory models the record linkage problem in practice. Applying the identification codes to the two-party privacy-preserving record linkage problem, we demonstrate the advantage on the performance analysis over the classical hash-based approaches, especially on the privacy analysis. Note for the PPRL, linkage quality is typically evaluated experimentally, and for privacy, there is so far no commonly accepted privacy measures available that allow an objective evaluation. Our approach of identification code provides an objective evaluation on both linkage quality and privacy based on parameters of identification codes.

Acknowledgment

The author would like to acknowledge the fruitful discussions with Prof. Rainer Schnell, Prof. Frederik Armknecht and Youthe Heng on PPRL.

References

1. Ahlswede, R., Dueck, G.: Identification via channels. IEEE Transactions on Information Theory **35**(1), 15–29 (1989)
2. Becker, G., Jenkin, B.: fastdigest: Fast, low memory-footprint digests of R objects. `https://CRAN.R-project.org/package=fastdigest`, R package version 0.6-3
3. den Boer, B.: A simple and key-economical unconditional authentication scheme. Journal of Computer Security **2**, 65–71 (1993)
4. Christen, P.: Data Matching - Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection. Springer, Berlin (2012)
5. Christen, P., Ranbaduge, T., Schnell, R.: Linking Sensitive Data - Methods and Techniques for Practical Privacy-Preserving Information Sharing. Springer, Berlin (2020)
6. Christen, P., Verykios, V.: A tutorial on privacy-preserving record linkage. `https://www.academia.edu/75300246/A_Tutorial_on_Privacy_Preserving_Record_Linkage`
7. Han, T., Verdu, S.: New results in the theory of identification via channels. IEEE Transactions on Information Theory **38**(1), 14–25 (1992)
8. Kurosawa, K., Yoshida, T.: Strongly universal hashing and identification codes via channels. IEEE Transactions on Information Theory **45**(6), 2091–2095 (1999)
9. Schnell, R., Bachteler, T., Reiher, J.: Privacy-preserving record linkage using Bloom filters. BMC Medical Informatics & Decision Making **9**, 41 (2009)
10. Shannon, C.E.: A mathematical theory of communication. Bell Syst. Tech. J. **27**(3), 379–423 (1948)
11. Vatsalan, D., Christen, P.: Scalable privacy-preserving record linkage for multiple databases (November 2014). https://doi.org/10.1145/2661829.2661875
12. Vatsalan, D., Christen, P., Verykios, V.: An efficient two-party protocol for approximate matching in private record linkage. vol. 121, pp. 125–136 (January 2011)
13. Verdu, S., Wei, V.: Explicit construction of optimal constant-weight codes for identification via channels. IEEE Transactions on Information Theory **39**(1), 30–36 (1993)

An Automated Approach to Collecting and Labeling Time Series Data for Event Detection Using Elastic Node Hardware

Tianheng Ling, Islam Mansour, Chao Qian, and Gregor Schiele

Intelligent Embedded Systems Lab, University of Duisburg-Essen,
47057, Duisburg, Germany
{tianheng.ling, chao.qian, gregor.schiele}@uni-due.de
islam.mansour@stud.uni-due.de

Abstract. Recent advancements in IoT technologies have underscored the importance of using sensor data to understand environmental contexts effectively. This paper introduces a novel embedded system designed to autonomously label sensor data directly on IoT devices, thereby enhancing the efficiency of data collection methods. We present an integrated hardware and software solution equipped with specialized labeling sensors that streamline the capture and labeling of diverse types of sensor data. By implementing local processing with lightweight labeling methods, our system minimizes the need for extensive data transmission and reduces dependence on external resources. Experimental validation with collected data and a Convolutional Neural Network (CNN) model achieved a high classification accuracy of up to 91.67%, as confirmed through 4-fold cross-validation. These results demonstrate the system's robust capability to collect audio and vibration data with correct labels.

Keywords: Event Detection · Time Series · Sensor Data Collection · Automated Labeling · Embedded Systems · CNN · Integrated Hardware System

1 Introduction

Event detection has become a popular topic in pervasive computing [1], enabling intelligent systems to interpret environmental contexts and adapt configurations within various spaces, for example, offices or kitchens [2, 3]. Traditional IoT methods often utilize multiple types of indirect sensor data, such as audio and vibrations [4], which are processed through Deep Learning (DL) models for event recognition.

Sufficiently labeled datasets are necessary to train DL models effectively [5]. Typically, data streams are segmented and annotated with labels [6]. One common approach to collecting these datasets involves transmitting sensor data to the cloud [7], where labeling algorithms are applied [8], or storing the data streams for subsequent manual labeling by human workers [9]. Both methods, however, introduce significant delays and dependencies on external resources.

Instead of transmitting data while collecting them, we propose a local processing approach. Given that IoT devices generally possess limited processing power [10], applying complex labeling algorithms in real-time during data collection poses significant challenges. To overcome this obstacle, we have developed a novel embedded system designed to collect and automatically label data using light-weight methods. This approach significantly reduces the need for continuous data transmission, aligning with the constraints of power, energy, and latency typical in the IoT context. The main contributions of this research include:

- We designed an integrated hardware system equipped with various sensors and an SD card slot to facilitate on-device data and label storage. We also included additional labeling sensors to ensure accurate and efficient event detection.
- We developed software that features a predefined set of labels. The labeling process is automated through an interrupt- and threshold-based detection mechanism, significantly simplifying the computation required for label extraction.
- We validated the efficacy of our collected dataset through experiments with event classification using a Convolutional Neural Network (CNN) model. On our custom dataset across three event types, our model achieved up to 91.67% test accuracy, verified through 4-fold cross-validation.

The remainder of this paper is organized as follows: Section 2 reviews relevant literature, setting the stage for our research. Section 3 details our hardware design, while Section 4 discusses the software implementation. Section 5 presents experiment setups and analyzes our findings. Finally, Section 6 concludes the paper and outlines directions for future research.

2 Related Work

Previous studies predominantly relied on human involvement in the recording and labeling process, which not only complicates the procedure but also increases costs and the potential for errors during manual operations.

Specifically, Koch et al. [11] manually controlled the start and stop of recordings for each event. While this method minimizes storage requirements, it introduces complexity and heightens the risk of human error. In contrast, Anand et al. [12] implemented continuous data recording with post-collection labels based on timestamps and event types. This method simplifies the recording process but often accumulates large volumes of irrelevant data, leading to inefficient storage usage, especially when events are infrequent. Furthermore, while humans can feasibly label audio data by listening, this approach is impractical for vibration data.

In response to these challenges, our research introduces a novel automated system that significantly reduces the need for manual intervention by automating the collection of reference labels. Our approach utilizes additional sensors that only need light-weight computation to determine the event type locally. With

an on-device approach, we are free from synchronization challenges and can efficiently capture the essential sensor data before and after an event occurs.

3 Hardware System Design

In this study, we propose a hardware platform named Elastic Node Sensor Logger. Figure 1 illustrates the architecture of our hardware platform, centered around the RP2040, an ARM Cortex-M0+ Microcontroller Unit (MCU) known for its low power consumption. In addition, its performance is sufficient for running FreeRTOS to make our multi-task scheduling easier than just using a bare-metal setup. This MCU also owns enough analog and digital I/O capabilities, which are crucial for managing the various sensors and storage modules incorporated into our system.

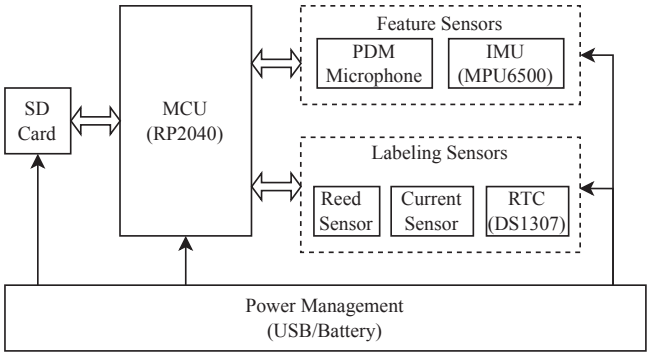


Fig. 1: System Architecture of Elastic Node Sensor Logger

Our system includes two categories of sensors: feature and labeling sensors. The feature sensors are supposed to monitor the events indirectly. There is a Pulse Density Modulation (PDM) microphone for collecting audio data, and an Inertial Measurement Unit (IMU) programmed as an accelerator meter for collecting vibration data. Their sampling frequency is a parameter that the user can adjust. This configuration facilitates the generation of time series data, which is vital for training our DL models.

For event labeling, we utilize a reed sensor and a current sensor. The reed sensor detects door states by issuing a rising edge interrupt to the MCU when the door opens and a falling edge interrupt upon closing. Concurrently, the current sensor monitors the power consumption of a kettle. We use the analog-to-digital converter on MCU to detect the 'water has boiled' event based on a predefined current threshold (zero). In addition, a lower-power Real Time Clock (RTC) is embedded in the board to provide the timestamp for events.

Additionally, an SD Card, connected to the MCU via the Serial Peripheral Interface, has been configured to operate at a writing speed with a clock fre-

quency of up to 50 MHz. This high speed far exceeds the data acquisition rates from all sensors, ensuring that data logging remains efficient and does not hinder the system's overall performance.

The power management subsystem, including the MCP73833 for battery charging and the LM1117-3.3 for voltage regulation, ensures sufficient power utilization across all components. Given that the peak current consumption is estimated at 440 mW, a low-dropout regulator, LM1117-3.3, is sufficiently adequate for our power regulation needs, promoting system stability and efficiency.

4 Software Implementation

The main software loop, executed on the MCU, is depicted in Figure 2. At the outset, we initialize the sensor drivers and mount the SD card, setting the stage for data collection.

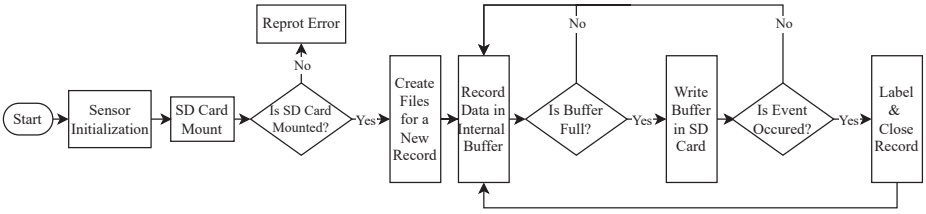


Fig. 2: Main Loop of the Recording

To optimize memory management and processing efficiency, we implement a ping-pong buffer strategy on the MCU for handling sensor data, a method akin to that described in [13]. Data is initially collected in the 'ping' buffer until it reaches capacity. At this point, data storage is switched to the 'pong' buffer. This cycle alternates to ensure continuous data acquisition. Upon filling either buffer, a Direct Memory Access (DMA) is triggered to transfer the data to the SD card, thereby offloading the data writing task from the MCU. This setup ensures that the SD card's write speed surpasses our data acquisition rate and provides ample buffer time to prevent overwriting and maintain data integrity.

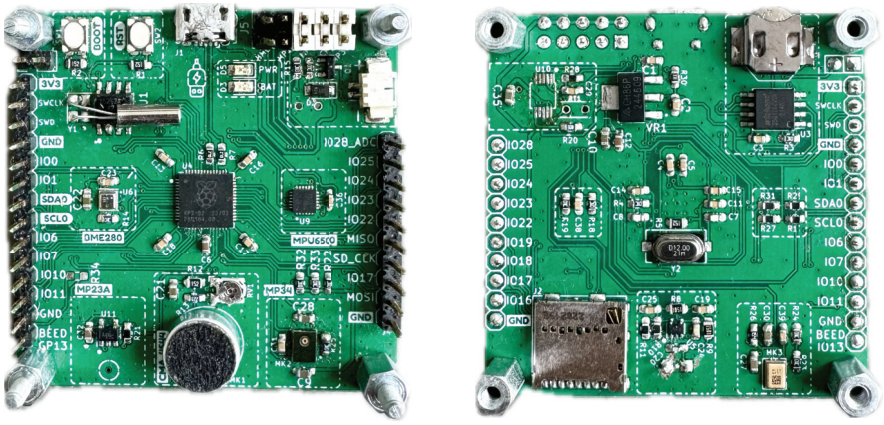
The system has three event flags, two set by external interrupt callbacks and the third by a threshold-based trigger following ADC readings in a separate periodic task. This event detection logic is straightforward and computationally efficient, avoiding disruptions in data collection.

The system checks for flagged events once the DMA completes the data transfer from one buffer. If an event has been flagged, the corresponding label is immediately written to the SD card. The recording file may be closed promptly or left open for several seconds to capture additional post-event data, and the duration of the continuous recording is user-configurable.

Furthermore, our system utilizes an *FatFs*[1] file system to support up to four simultaneous file operations on the SD card. This capability allows for concurrent audio and vibration data recording, each stored in formats optimized for ease of access and analysis. Audio recordings are saved in WAV format for convenient review, while vibration data and event labels with timestamps are stored in separate CSV files, simplifying data management and enhancing accessibility.

5 Experiments and Results

Building upon the hardware system design outlined in Section 3, we successfully implemented the hardware platform as depicted in Figure 3. Utilizing this hardware and following the software implementation described in Section 4, we conducted multi-sensor data collection and labeling directly on our hardware platform. Subsequently, the collected dataset underwent preprocessing and validation on a desktop computer.



(a) Front Side (b) Back Side

Fig. 3: Elastic Node Sensor Logger

5.1 Multi-Sensor Data Collection

The data collection process solely requires the use of the Elastic Node Sensor Logger hardware. As mentioned in Section 4, once the recording process initiates, our system simultaneously collects data from the microphone and the IMU sensor. The labeling sensors operate in the background as described. The audio data is captured at a sampling rate of 16 kHz in a mono-channel format. The

[1] https://github.com/elehobica/pico_fatfs

vibration data, which includes three channels corresponding to acceleration, is collected at a sampling rate of 4 kHz. After starting the recording, the device operates autonomously for several hours. During this period, the user (operator) randomly engages in activities such as opening and closing doors and boiling water to generate event data. In total, we collected 106 samples from each type of sensor: 40 samples were associated with door opening events, 29 with door closing, and 37 with water boiling in a kettle.

5.2 Data Preprocessing

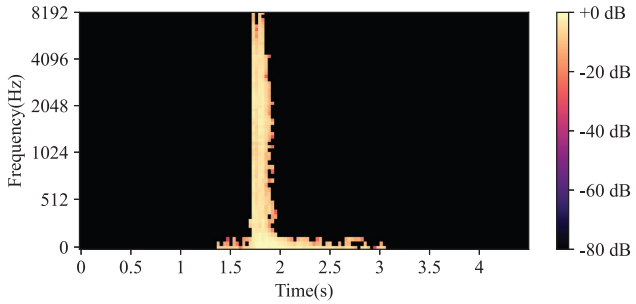
Before we fed our custom dataset to the model, audio and vibration data were preprocessed separately to accommodate their unique characteristics . Audio recordings were read from WAV files using torchaudio [14]. To standardize the lengths of these recordings, zero-padding was symmetrically applied to both ends to have the same length as the longest audio data in the dataset. For feature extraction, we transformed the recordings into Mel spectrograms using the following parameters: $n_{\text{mels}} = 64$, $n_{\text{fft}} = 1024$, with hop_{length} at default settings, and $top_{\text{db}} = 80$ for dynamic range compression. Figure 4 displays these spectrograms for three events, showcasing their distinct spectral characteristics. For example, door-related events exhibit short-term peaks along the time axis and a broader range of frequency coverage compared to kettle-boiling events. Furthermore, distinguishable patterns are evident between door opening and closing events, such as the longer duration of closing events compared to opening events.

Vibration recordings, comprised of channel measurements, varied in length across samples. We addressed this by applying a zero-padding strategy similar to that used for audio data. In instances of missing values, we imputed these by calculating and using the mean of the respective dimension. Figure 5 displays the visualization of vibration data, categorized in terms of the three sample events. Distinct patterns are identifiable between the door-related events, and there is a clear difference between these and the water-boiling events in the time domain.

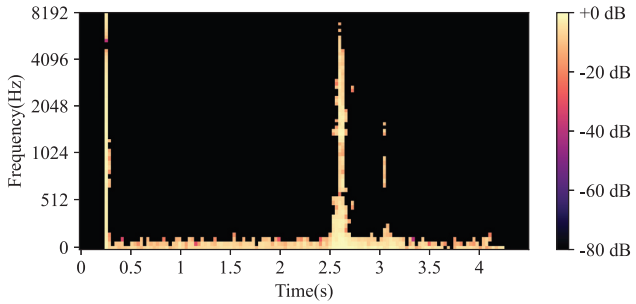
5.3 Data Validation

To verify the quality of the collected data and the accuracy of labeling, we conducted a three-class classification task based on audio and vibration data. We consider the quality of our collected dataset to be high if the collected data and labels can train a deep learning model to converge and achieve test high accuracy. We split the entire dataset into training, validation, and testing sets in a 3:1:1 ratio. Afterward, we utilized an oversampling strategy to balance the distribution of samples across different labels for both data types. Training and validation sets underwent a 4-fold cross-validation process. Moreover, we computed the mean and standard deviation based on the training set to normalize all datasets.

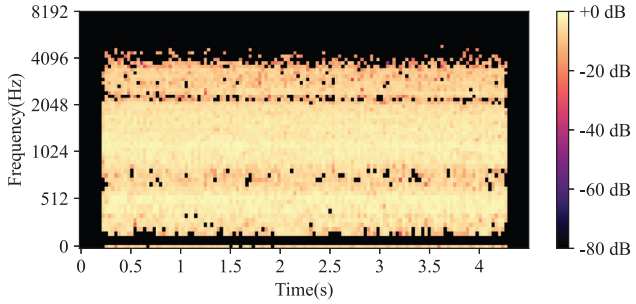
We adopted a simple CNN model for event classification under the PyTorch framework. As depicted in Figure 6, this model features two convolutional layers, with the first layer having 64 filters and the second 32 filters using a kernel



(a) Close Door



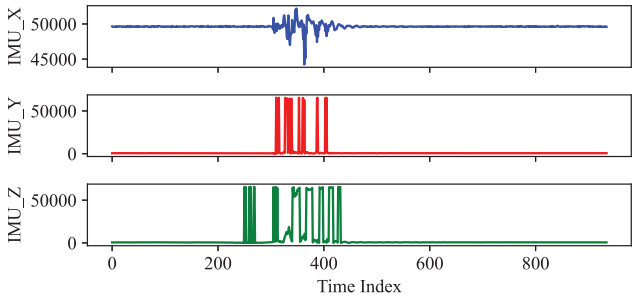
(b) Open Door



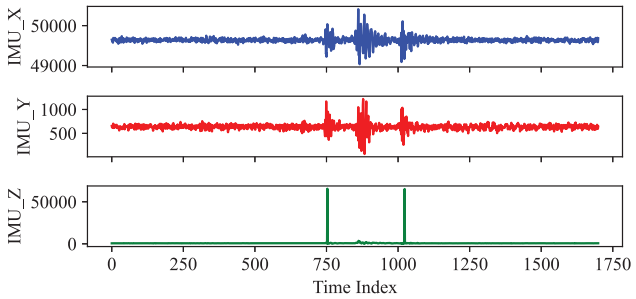
(c) Water Boiling

Fig. 4: Mel Spectrograms of Different Event Labels

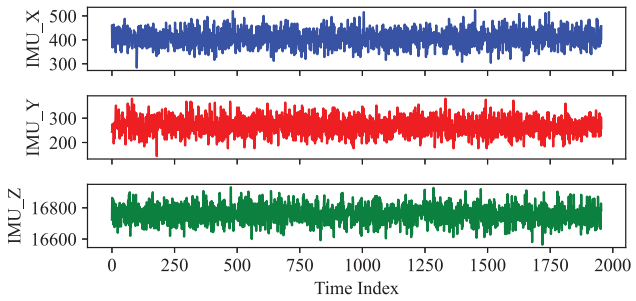
size of 3 and stride of 1. Each convolutional layer is equipped with Rectified Linear Unit (ReLU) activation function and batch normalization, enhancing the model's learning efficiency. They are then followed by an adaptive average pooling layer that reduces dimensionality, preparing the output for the final classifi-



(a) Close Door



(b) Open Door



(c) Water Boiling

Fig. 5: Visualization of vibration data with different event labels

cation stage. The processed data is fed into a fully connected layer, classifying events.

We configured our model training using the *Adam* optimizer, setting hyperparameters to $\beta_1 = 0.9$, $\beta_2 = 0.98$, and $\epsilon = 10^{-9}$. The training initiated with a learning rate of 0.001, which we dynamically adjusted using a scheduler that modified the rate at a step size of 3 with a decay factor γ of 0.5. We opted for cross-entropy error as the loss function to train and evaluate the model's

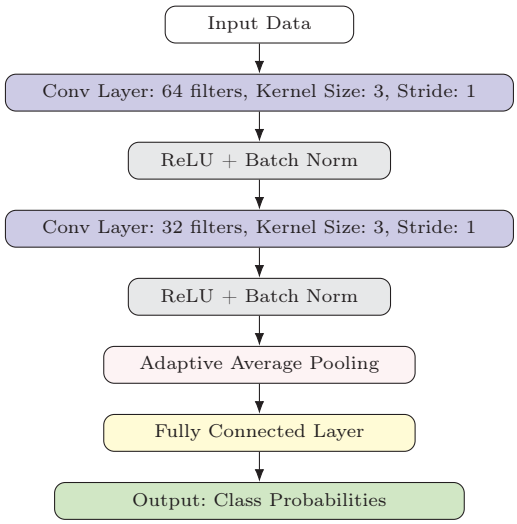


Fig. 6: CNN Architecture for Time-series Data Classification

performance. To enhance the robustness of our training process, we conducted 100 experiments, each comprising 50 epochs, and incorporated an early stopping mechanism to mitigate the risk of overfitting. We used accuracy as the primary evaluation metric complemented by a confusion matrix to provide detailed insights into the model's performance across different events.

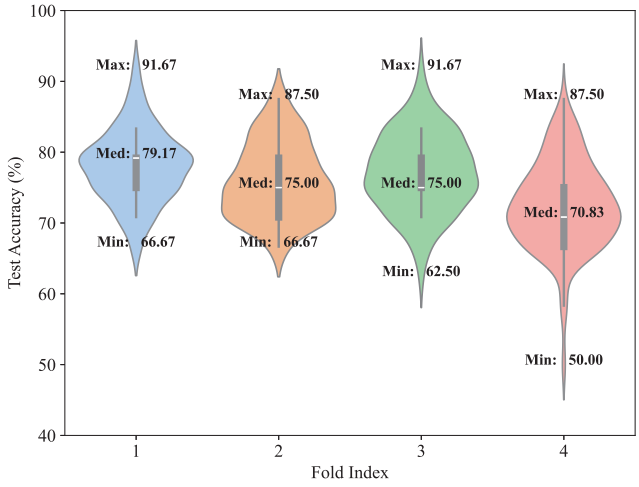


Fig. 7: Audio Data: Test Accuracy (%) across Different Folds

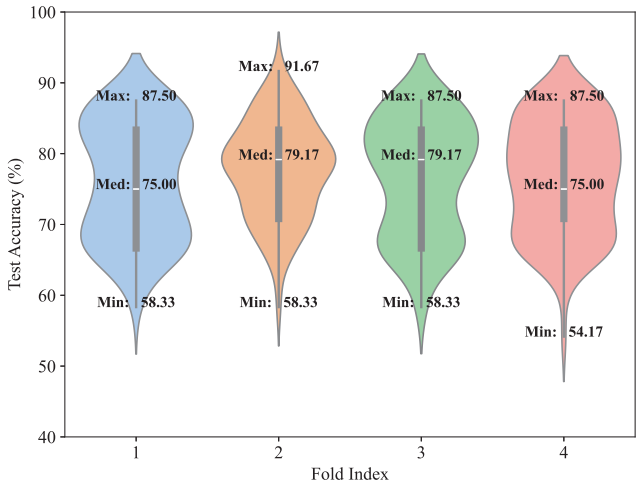


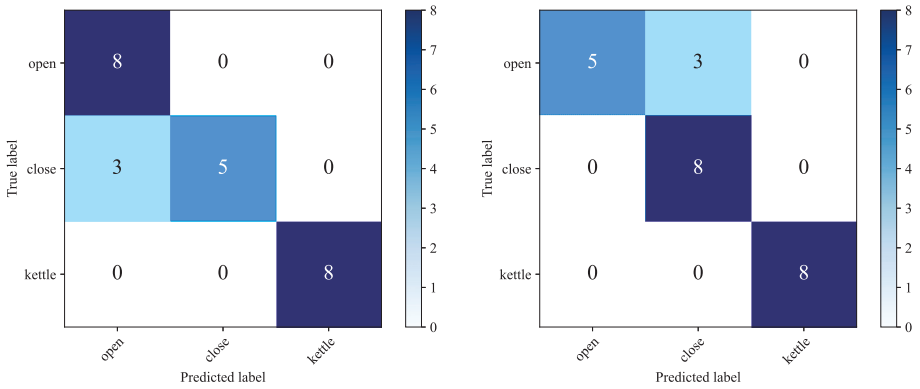
Fig. 8: Vibration Data: Test Accuracy (%) across Different Folds

Figure 7 illustrates the distribution of test accuracy for audio data across different validation folds. The observed minimum accuracy ranged from 50.00% to 66.67%. Despite these variations, the model demonstrates strong potential, achieving maximum accuracy up to 91.67% in folds 1 and 3, and 87.50% in folds 2 and 4. The median accuracy, spanning from 70.83% to 79.17%, suggests that the model generally maintains high-performance levels. Figure 8 presents the test accuracy for vibration data, which also exhibits variability with minimum accuracy between 54.17% and 58.33%. The model reaches a high accuracy of up to 91.67% in fold 2 and consistently above 87.50% in the other folds. The median accuracy, consistently between 75.00% and 79.17%, indicates a reliable performance.

To understand the limitations of our system and identify potential areas for improvement, we conducted a detailed analysis of a trained model with a test accuracy of 87.5% using a confusion matrix. Figure 9 (a) displays the confusion matrix for the model trained with audio data. It reveals that all samples of water boiling and door opening are correctly classified, although three samples of door closing were misclassified as door opening. Similarly, Figure 9 (b), which pertains to the model trained with vibration data, shows that only three samples of door-opening events were misclassified as door-closing events.

It is important to note that although we collected audio and vibration data concurrently, the models were trained separately on each data type. Integrating both audio and vibration data as inputs for the models could potentially enhance accuracy further, particularly in applications requiring high precision. However, investigating this integrated approach is beyond the scope of this paper.

In summary, the quality of the collected data and labels has proven sufficient for CNN to learn and differentiate between event classes effectively. While the



(a) Audio Data (b) Vibration Data

Fig. 9: Confusion Matrix With Test Accuracy 87.5%

classification accuracy from vibration data is slightly lower than that from audio data, this outcome was anticipated due to the inherent challenges associated with vibration signal classification. Notably, the consistency of our sensor data has been validated across four-folds, confirming the effectiveness of our system in capturing classifiable features across three distinct event types.

6 Conclusion and Future Work

Our study successfully developed a robust approach for autonomously labeling sensor data directly on IoT devices. Experiments demonstrated that our models achieved up to 91.67% test accuracy in controlled settings, highlighting the high quality of our sensor data and the reliability of our labeling approach. This method significantly improves the feasibility of collecting and processing large-scale IoT data in diverse field environments, enhancing efficiency and accuracy.

In our ongoing efforts to enhance event detection capabilities, we plan to integrate additional types of feature sensors into our system. This expansion will enable the support and recognition of a broader array of event types, further improving the versatility and applicability of our solution in diverse scenarios. By broadening the sensor array, we aim to capture more comprehensive feature data from the device surroundings, significantly refining our system's responsiveness and accuracy in real-world applications.

Acknowledgments

The authors gratefully acknowledge the financial support provided by the Federal Ministry for Economic Affairs and Climate Action of Germany for the RIWWER project (01MD22007C).

References

1. Yu, M., Bambacus, M., Cervone, G., Clarke, K., Duffy, D., Huang, Q., Li, J., Li, W., Li, Z., Liu, Q., et al.: Spatiotemporal event detection: a review. International Journal of Digital Earth **13**(12), 1339–1365 (2020)
2. Vafeiadis, A., Votis, K., Giakoumis, D., Tzovaras, D., Chen, L., Hamzaoui, R.: Audio content analysis for unobtrusive event detection in smart homes. Engineering Applications of Artificial Intelligence **89**, 103226 (2020)
3. Pandya, S., Ghayvat, H.: Ambient acoustic event assistive framework for identification, detection, and recognition of unknown acoustic events of a residence. Advanced Engineering Informatics **47**, 101238 (2021)
4. Choudhary, P., Kumari, P., Goel, N., Saini, M.: An audio-seismic fusion framework for human activity recognition in an outdoor environment. IEEE Sensors Journal **22**(23), 22817–22827 (2022)
5. Alzubaidi, L., Bai, J., Al-Sabaawi, A., Santamaría, J., Albahri, A.S., Al-dabbagh, B.S.N., Fadhel, M.A., Manoufali, M., Zhang, J., Al-Timemy, A.H., et al.: A survey on deep learning tools dealing with data scarcity: definitions, challenges, solutions, tips, and applications. Journal of Big Data **10**(1), 46 (2023)
6. Bouchabou, D., Nguyen, S.M., Lohr, C., LeDuc, B., Kanellos, I.: A survey of human activity recognition in smart homes based on IoT sensors algorithms: taxonomies, challenges, and opportunities with deep learning. Sensors **21**(18), 6037 (2021)
7. Ghosh, A.M., Grolinger, K.: Edge-cloud computing for internet of things data analytics: embedding intelligence in the edge with deep learning. IEEE Transactions on Industrial Informatics **17**(3), 2191–2200 (2020)
8. Pičuljan, N., Car, Ž.: Machine learning-based label quality assurance for object detection projects in requirements engineering. Applied Sciences **13**(10), 6234 (2023)
9. Fahy, C., Yang, S., Gongora, M.: Scarcity of labels in non-stationary data streams: a survey. ACM Computing Surveys (CSUR) **55**(2), 1–39 (2022)
10. Qian, C., Ling, T., Schiele, G.: Enhancing energy-efficiency by solving the throughput bottleneck of LSTM cells for embedded FPGAs. In: Joint European Conference on Machine Learning and Knowledge Discovery in Databases. pp. 594–605. Springer (2022)
11. Koch, M., Schlenke, F., Kohlmorgen, F., Kuller, M., Bauer, J., Woehrle, H.: Detection and classification of human activities using distributed sensing of environmental vibrations. In: 2022 IEEE International Conference on Omni-layer Intelligent Systems (COINS). pp. 1–6. IEEE (2022)
12. Anand, J., Koch, M., Schlenke, F., Kohlmorgen, F., Wöhrle, H.: Classification of human indoor activities with resource constrained network architectures on audio data. In: 2022 IEEE 5th International Conference and Workshop Óbuda on Electrical and Power Engineering (CANDO-EPE). pp. 000157–000162. IEEE (2022)
13. Prasad, M.: Ping-pong buffers. https://onlinedocs.microchip.com/pr/GUID-324A966D-1464-4B35-A7D1-DCAE052AC22C-en-US-3/index.html?GUID-B6995A5F-E06B-4071-893E-BBC60082F576 (2024)
14. Yang, Y.Y., Hira, M., Ni, Z., Astafurov, A., Chen, C., Puhrsch, C., Pollack, D., Genzel, D., Greenberg, D., Yang, E.Z., et al.: Torchaudio: building blocks for audio and speech processing. In: ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 6982–6986. IEEE (2022)

Towards Training DNNs with Quantized Parameters

Leo Buron[1,2][0009−0001−8939−4784], Lukas Einhaus[1,2][0000−0002−6102−7077], Andreas Erbslöh[1,2][0000−0001−6702−892X], and Gregor Schiele[1,2][0000−0003−4266−4828]

[1] University of Duisburg-Essen, LAB for Intelligent Embedded Systems, Forsthausweg 2, 47057 Duisburg, Germany firstname.lastname@uni-due.de https://www.uni-due.com/es/

[2] paluno - The Ruhr Institute for Software Technology, University of Duisburg-Essen, Essen, Germany https://paluno.uni-due.de/

Abstract. For embedded devices, memory and computational cost are limited. However, the training of neural networks is computation and memory expensive. To reduce memory consumption quantization schemes are applied on the parameters by a lot of related work. The introduced artifacts can be reduced by applying stochastic rounding instead of round-half-to-even to the quantization scheme. While a lot of related work is using stochastic rounding for inference and gradient computation, none have shown the benefit of it for the parameter update only. We use a fixed point quantization scheme and propose to quantize parameters at model initialization. When training, we use full-resolution gradient computation and apply stochastic quantized updates. The quantization of the parameters and the parameter update lead to reduced memory consumption. In addition, the inference is mostly quantized, speeding up computations. For training, the inputs of each layer are stored for gradient computations. Due to quantized inference those inputs are already quantized reducing the needed memory further. We explore a model for different fixed point configurations on the FASHION-MNIST dataset. Generally, stochastic rounding parameter updates match or beat the top-5 accuracy of QAT. We are able to achieve 0.876 top-5 accuracy while reducing the memory to 0.93. When using stochastic rounding as well for inference, we can achieve 0.737 top-5 accuracy for a memory reduction of 0.88.

Keywords: deep learning · quantized parameters · edge training · embedded devices · memory reduction · computational cost · stochastic round-ing · inference · fixed point quantization

1 Introduction

With the recent success of deep neural networks, more challenging tasks are solved with more complex and bigger neural networks. The increase in network parameters makes the training challenging even on GPU clusters, where memory is often the bottleneck to fit these models on GPUs [5]. This is a more

challenging task for edge computing and embedded computing, where computing resources are typically limited by design. At the same time, the demand for local computing on these resource-constrained devices is increasing. Applications like brain-computer-interfaces [4,7,8] and human activity recognition with personal devices like smartphones and smartwatches [10,15] raise the need for data privacy, network independence and reduced transmission energy, and low latency not only for local inference, but also training of the neural network.

For inference and training of neural networks usually floating point arithmetic using 32 bits (*FlP32*) is used in calculations. To reduce the bit width of parameters quantization schemes can be applied. However, the direct application of quantization schemes can reduce the network's performance metric. Therefore, well-known techniques, like quantization-aware training (QAT), reduce the performance loss when using quantized parameters' for the inference. However, they do not reduce the parameters' bit width during training. A well known quantization scheme that also computes faster on embedded devices is the FxP quantization scheme, consisting of an integer part and a fractional part. The FxP quantization scheme is as follows: First, the values are clamped to its two's complement value range. Then the value is rounded to its last fractional digit. The rounding mode round half-to-even (RHE) is typically used because scientific rounding would always round up at half increasing the result and accumulating to a bigger error. When multiplying two values with a fractional part the size of the fractional part doubles. Thus, the result is rounded to the original fractional length. One challenge is the loss of precision when rounding. A well-known solution, called stochastic rounding (SR) [6,12,16,18,19,3], samples from a uniform distribution and adds it on the input prior to rounding.

In this study, we show that the memory consumption during training can be reduced with fixed point (FxP) quantization scheme for parameters, layer outputs and momentum in stochastic gradient descent (SGD) optimizer. We demonstrate, that the use of stochastic rounding (SR) solely in the optimizer achieves comparable performance to state-of-the-art ($SOTA$) methods like quantization-aware training (QAT). We test this for different FxP quantization scheme configuration on the FASHION-MNIST dataset [17]. In addition, we show that the use of stochastic rounding in the forward pass can regain some loss for lower bit-width.

Our main contributions are as follows:

- First, we present an approach for the SGD-based optimizer allowing quantized parameter updates.
- We show how the memory consumption for parameters and optimizer momentum reduces compared to FlP32.
- Also, we show that SR during the inference can additionally benefit the accuracy.

In the next chapter, we present the related work. The following chapter describes our method. We show the experiments of our paper in the fourth chapter. Finally, we describe our findings and present our future work.

2 Related Work

The related work is split into two fields. One field is training to get a quantized model. The other one is training a quantized model. There are two SOTA methods, one beeing quantization-aware training and post-training quantization (PTQ). While post training quantization does not reduce any memory consumption during training, we do not consider PTQ further. With quantization-aware training (QAT) the goal is to reduce the introduced quantization error by applying a quantization during training. This allows the optimizer to reduce its loss. One technique called straight-through estimator[1] uses a full-resolution copy of the parameters. All parameters are quantized during inference. The inputs of each layer are quantized reducing the needed memory for the gradient calculation.

More recent research is also training with quantized parameters. All of those papers are quantizing both forward and backward pass to reduce memory and computation costs. [11] are using half precision (16bit) floating point IEEE for all multiplications, but also use FlP32 accumulator. In addition, they use a FlP32 copy of their parameters. [6] uses 16 bit FxP quantization for all computations. Also, they identify *SR* as a key technique for the backward pass. [16] also shows that for 8 bit floating point format using four exponent bits in inference and five exponent bits in the gradient. They use *SR* for their calculations. [12,3] both use a block floating point format, sharing one exponent for a tensor reducing the memory consumption. Their implementation is also using SR. [19] implement low-bit quantization scheme with SR. The briefly presented papers [6,16,12,19,3] quantize the backward pass which leads to higher loss. They state that they apply *SR* on all calculations. We believe they also apply *SR* to the calculations in the optimizer. However, those are never described explicitly. To the best of our knowledge, no paper has examined the influence of *SR* limited to the optimizers' calculations for quantized parameters. Therefore, we quantize the inference, but not the gradient calculations.

3 Method

We describe the FxP quantization scheme with a FxP configuration using three parameters; total number of bits $TotalBits$, fractional number of bits $FracBits$, and rounding mode either RHE or SR. We implement the quantization depending on the rounding mode. The implementation for rounding mode RHE is shown in equation 1. For stochastic rounding a noise is sampled from a uniform distribution $\mathcal{U}$ and scaled depending on the number of fractional bits $FracBits$ (see equation 2). The noise is added on the input before applying the quantization 3.

$$quantizeRHE(x) = \hat{x} = \frac{RHE(x * 2^{FracBits})}{2^{FracBits}} \tag{1}$$

$$noise = \frac{Sample_{\mathcal{U}} - 0.5}{2^{FracBits}} \quad \text{with } \mathcal{U}(0,1) \tag{2}$$

$$quantizeSR(x) = quantize(x + noise) \tag{3}$$

During the initialization of the neural network, we apply the FxP quantization scheme with the chosen configuration on each parameter. Thus, applying this quantization scheme reduces the values bitwidth. We describe the memory reduction as in equation 4. In this study, we compare the memory consumption against FlP32. This means that a FxP value with 16 total bits has a memory reduction of 0.5.

$$\begin{aligned} MemoryReduction &= \frac{MemoryConsumption(\hat{x})}{MemoryConsumption(x)} \\ &= \frac{TotalBits(\hat{x})}{TotalBits(x)} \end{aligned} \tag{4}$$

During inference, the same quantization scheme is applied for all layer outputs. For the calculation of the gradients, each layer's input needs to be saved during inference. The optimizer is calculating the parameter update $\delta_{w,t}$ for each parameter w at time point t. When using stochastic gradient descent optimizer (SGD) $\delta_{w,t}$ is computed with the gradient $g_{w,t}$, and the learning rate γ is applied as a scaling factor. In SGD a momentum is added on the gradient $g_{w,t}$. The momentum is calculated with the momentum buffer $\delta_{w,t-1}$, which is down scaled by the momentum factor μ (see equation 5).

$$\delta_{w,t} = (\mu * \delta_{w,t-1} + g_{w,t}) * \gamma \tag{5}$$

Each parameter w is updated by substracting $\delta_{w,t}$ from w_t resulting in w_{t+1}(see equation 6). To update the quantized parameters accordingly, the FxP quantization scheme needs to be applied on $\delta_{w,t}$. The quantization of δ is important for the update of $\delta_{w,t}$, but is also stored in the momentum buffer reducing its memory consumption as described in equation 4.

$$w_{t+1} = w_t - \delta_{w,t} \tag{6}$$

This technique comes with the overhead of SR during the parameter update. Since we do not specify a target platform the cost for the technique can vary. We describe the cost C for SR C_{SR} with the cost for sampling from a uniform distribution $C_{Sample_{\mathcal{U}}}$, the cost of a bit shift $C_{BitShift}$ for number of fractional bits, the cost for adding C_{Add} that on the input signal, and the cost for apply RHE C_{RHE} in equation 7. When using SR in inference, costs are incurred for each rounding. For QAT, this overhead increases more for inference because each parameter is quantized during inference.

$$C_{SR} = C_{Sample_{\mathcal{U}}} + C_{BitShift} * FracBits + C_{Add} + C_{RHE} \tag{7}$$

We hypothesise that the use of a **s**tochastic **r**ounding for the **p**arameter **u**pdate ($SRPU$) with quantized $\delta w, t$ and following with quantized momentum

buffer $\delta_{w,t-1}$ achieves comparable performance for the same quantization scheme when using quantization-aware training (QAT) method. We test our hypothesis for different FxP configurations for parameters and inference calculations with both **s**tochastic **r**ounding for the **p**arameter **u**pdate ($SRPU$) and for **r**ound **h**alf-to-**e**ven for the **p**arameter **u**pdate ($RHEPU$). In addition, we apply the same FxP configurations with the use of QAT-method.

4 Experiments

In this section, we test whether our hypothesis stands for a set of FxP configurations applied on one model architecture for the FASHION-MNIST dataset [17]. The dataset consists of ten categories of fashion clothes represented in 8-bit gray-scale images. The dataset is split into 60000 training samples and 10000 test samples. The batch size of the training dataset is set to 256. The data is normalized to a mean of 0.1307 with a standard deviation of 0.3081. We do not quantize the input nor the data as we believe that related work is doing the same since they are not stating it explicitly.

Layer	Kernel	Channels	Filters	Feat-In	Feat-Out	Inputs	Round	Parameters
Conv2d	5x5	1	32	28x28	24x24	784	18432	800
MaxPool2d	2x2	32	32	24x24	12x12	18432	-	-
Conv2d	5x5	32	64	12x12	8x8	4608	4096	51200
MaxPool2d	2x2	64	64	8x8	4x4	4096	-	-
Linear	-	1	1	1024	128	1024	128	131200
Linear	-	1	1	128	10	128	10	1290
Softmax	-	1	1	10	10	10	-	-
					SUM	**29082**	**22666**	**184490**

Table 1. Model architecture derived from [6]

To evaluate the stochastic update, we train a convolutional neural network (see table 1), which is used in [6] for the MNIST dataset [2]. Each convolution is implemented without bias, no padding, a dilation of one, and a stride of one. Each linear layer uses a bias. Except for the last layer, each linear and convolutional layer is implemented with a ReLU activation function. The last layer uses the softmax activation function. We run the experiment once with in full-resolution (*FlP32*) to establish a baseline. Except for the baseline model, we quantize each learnable parameter, the output of each convolution, and linear layer with the same fixed point configuration. The output of the softmax layer is not quantized. We run each FxP configuration twice. Once with *RHE* and once with *SR* for the forward pass calculations. As loss function we chose the cross-entropy loss function. We use SGD-optimizer only with momentum. After a small grid search with the baseline model we set the SGD-optimizer's learning rate to 0.01 and the momentum to 0.9 confirming the choice of [6]. We

choose the same learning rate and momentum for all other training runs. We do not use a learning rate scheduler. We run each parameter update for each FxP configuration with the quantization-aware training (*QAT*) method, **s**tochastic **r**ounding for the **p**arameter **u**pdate (*SRPU*), and **r**ound **h**alf-to-**e**ven for the **p**arameter **u**pdate (*RHEPU*) for the same five seeds. Thus, we have no difference during model initialization and can compare the parameter update methods better. Each training run is 50 epochs long. The loss functions', gradients', and optimizers' calculations are never quantized. The experiment is implemented in PyTorch.

Next, we evaluate the memory consumption, then the computational overhead and finally the models Top-1 and Top-5 accuracy.

The size of the model parameters scales linearly with the number of total bits. For 32 bit the size of the model parameters is 737.96kB, for 8 bit it is 184.49kB. In figure 1 we show the needed memory for the five different total bit width. The memory reduction compared to *QAT* depends on the batch size. For a batch size of 256, the overall size of values to store is dominated by the inputs we need to store for the backward pass. For smaller batch sizes the bit width is more important. We take into account that the input for the neural network is not quantized, which leads to the affine behavior of the stored input per batch and therefore also of the memory reduction.

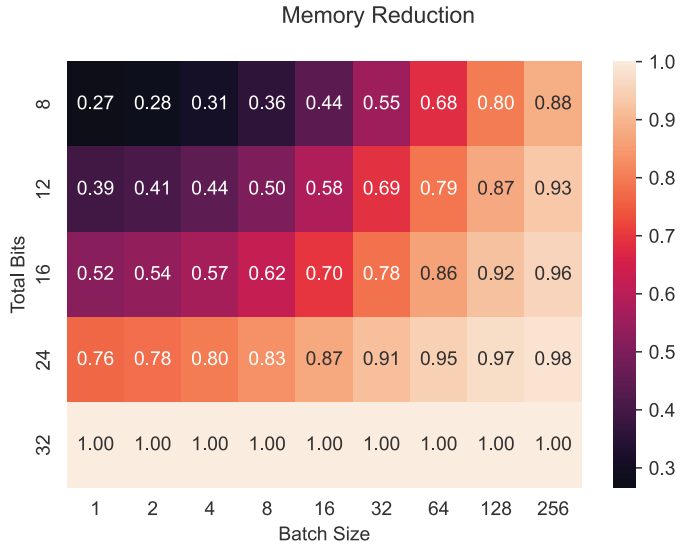


Fig. 1. Memory Reduction depending on total bit width and batch size compared to *QAT*

The overhead for the technique depends heavily on the used rounding mode during inference. *QAT* is rounding each parameter and layer output during in-

ference, while training with quantized parameters is only rounding layer outputs during the inference. In addition, training with quantized parameters needs to round for each parameter update. When using *SR* during inference, both *QAT* and *SR* parameter update methods have the same amount of rounding operations. Additionally, for *QAT* an up scaling prior to rounding and down scaling after rounding is applied during quantization.

Param Update	*QAT*	*SR*	*RHE*
Inference RM	207156	22666	22666
Param Update RM	0	184490	184490
Sum	207156	207156	207156

Table 2. Number of rounding operations depended on the rounding mode (RM)

After comparing the memory reduction and overhead, we are comparing the models Top-1 and Top-5 accuracies. The baseline model reaches a Top-1 accuracy of 0.755 and a Top-5 accuracy of 0.901 with a standard deviation of 0.064. We show a heatmap of the Top-1 and Top-5 validation accuracy for all tested FxP configuration with *RHE*during inference split by the parameter update method in figure 2. We show the same graph for the FxP configurations with *SR* during inference in figure 3.

First, we take a look at figure 2. Independent of the parameter update method the FxP configurations with 32 and 24 total bits perform equal to the baseline. For less than 16 total bits *RHEPU* methods performance drops to less than 0.12 for all runs. The Top-5 performance of *SRPU* always matches *QAT* performance. The performance of the baseline can not be matched by both approaches for 8 total bits. For 12 total bits 6 fractional bits match the performance of the baselines, but for 8 fractional bits the performance drops to 0.65. We believe that applying the same quantization scheme on both parameters and layer output is not optimal, because the accumulation of values can lead to higher values. Therefore, we propose to use a higher precision, but lower range for parameters and a lower precision, but higher range for layer outputs. This is also done by [6].

The figure 3 shows the Top-1 and Top-5 accuracy for the different FxP configurations dependent on the parameter update method with *SR* for inference. We compare this method, because the number of *SR* operations from *QAT* and *SR* are equal. *RHEPU* performs a little bit better in Top-5 accuracy for total bits lower or equal to 16, but never more than 0.18. All other accuracies for *RHEPU* look equal to the *RHE* inference mode. *SRPU* performance is almost equal for 16 total bits or more. *QAT* performance is equal for 24 and 32 total bits, but its Top-1 performance drops to 0.756 for 16 total bits. For 12 total bits and 8 fractional bits *SRPU* gaining 0.066 for Top-1 accuracy to a total of 0.814. For 8 total bits *QAT* and *SRPU* reach their highest accuracy for 3 fractional bits and gain up to 0.737. However, *SRPU* Top-1 accuracy can only reach 0.156. For 4 fractional bits *SRPU* Top-5 accuracy reaches 0.68, while *QAT* can only reach 0.215. For 5 total bits *QAT* loses performance compared to *RHE* during

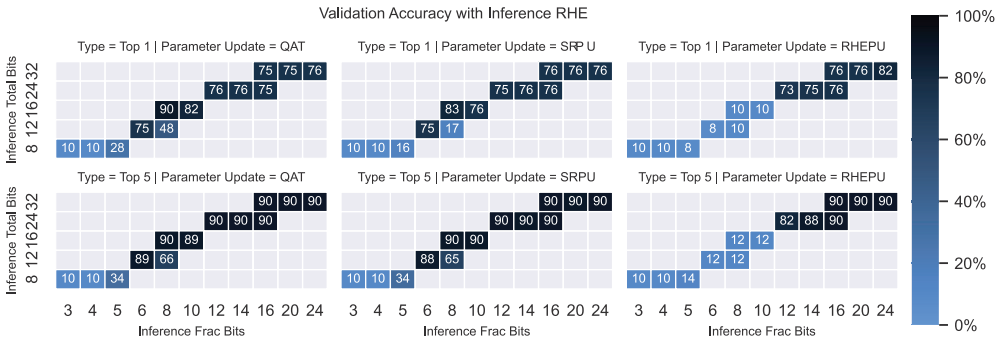


Fig. 2. Top-1 & Top-5 Validation Accuracy using *RHE*in inference for separated by the parameter update method

inference while *SRPU* remains equal. While *SRPU* can match or outbeat *QAT* performance generally, the difference in Top-1 and Top-5 accuracies varies a lot more for SRPU. This indicates that the training with *QAT* is more stable than *SRPU* for lower total bit widths.

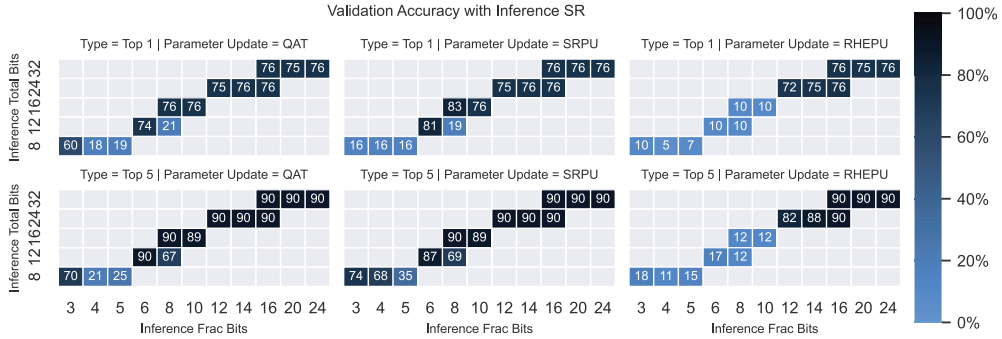


Fig. 3. Top-1 & Top-5 Validation Accuracy using *SR* in inference for separated by the parameter update method

5 Conclusion

We showed that the use of stochastic rounding during weight updates for quantized parameters can reduce memory consumption while achieving comparable performance to *QAT*. On Fashion-MNIST, we achieve 0.876 top-5 accuracy for 12 total bits with the use of round-half-to-even while reducing the used memory for

parameters and inputs during training to 0.93. For 8 total bits we could achieve 0.737 top-5 accuracy when using stochastic rounding also during inference.

In the future we plan to explore datasets like SVHN [13], CIFAR [9], and ImageNet [14]. Additionally, we want to improve the model by using different quantization scheme like block floating point to reduce the memory consumption further also tackling the problem that the weights should have a higher precision than layer outputs. Also, we want to fully quantize the gradients to reduce the computational cost. Finally, we want to evaluate those models on an FPGA.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Bengio, Y., Léonard, N., Courville, A.: Estimating or propagating gradients through stochastic neurons for conditional computation (2013). https://doi.org/10.48550/ARXIV.1308.3432
2. Bottou, L., Cortes, C., Denker, J., Drucker, H., Guyon, I., Jackel, L., LeCun, Y., Muller, U., Sackinger, E., Simard, P., Vapnik, V.: Comparison of classifier methods: a case study in handwritten digit recognition. In: Proceedings of the 12th IAPR International Conference on Pattern Recognition. IEEE Comput. Soc. Press (1994). https://doi.org/10.1109/icpr.1994.576879
3. Fox, S., Rasoulinezhad, S., Faraone, J., Boland, D., Leong, P.: A block minifloat representation for training deep neural networks. In: International Conference on Learning Representations (2021), https://openreview.net/forum?id=6zaTwpNSsQ2
4. Grahn, P., Mallory, G., Berry, M., Hachmann, J., Lobel, D., Lujan, L.: Restoration of motor function following spinal cord injury via optimal control of intraspinal microstimulation: toward a next generation closed-loop neural prosthesis. Frontiers in Neuroscience (2014). https://doi.org/10.3389/fnins.2014.00296
5. Guo, J., Liu, W., Wang, W., Yao, C., Han, J., Li, R., Lu, Y., Hu, S.: AccUDNN: A GPU memory efficient accelerator for training ultra-deep neural networks. In: 2019 IEEE 37th International Conference on Computer Design (ICCD). IEEE (2019). https://doi.org/10.1109/iccd46524.2019.00017
6. Gupta, S., Agrawal, A., Gopalakrishnan, K., Narayanan, P.: Deep learning with limited numerical precision. In: Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37. ICML'15, JMLR.org (2015)
7. Im, M., Kim, S.: Neurophysiological and medical considerations for better performing microelectronic retinal prothesis. Journal of Neural Engineering (2020). https://doi.org/10.1088/1741-2552/ab8ca9

8. Kathe, C., Skinnider, M., Hutson, T., et al.: The neurons that restore walking after paralysis. Nature (2022). https://doi.org/10.1038/s41586-022-05385-7
9. Krizhevsky, A.: Learning multiple layers of features from tiny images (2009), https://www.cs.toronto.edu/ kriz/cifar.html
10. Li, Y., Shi, D., Ding, B., Liu, D.: Unsupervised Feature Learning for Human Activity Recognition Using Smartphone Sensors. Springer International Publishing (2014). https://doi.org/10.1007/978-3-319-13817-6_11
11. Micikevicius, P., Narang, S., Alben, J., Diamos, G., Elsen, E., Garcia, D., Ginsburg, B., Houston, M., Kuchaiev, O., Venkatesh, G., Wu, H.: Mixed precision training. In: International Conference on Learning Representations (2018), https://openreview.net/forum?id=r1gs9JgRZ
12. Gennari do Nascimento, M., Adrian Prisacariu, V., Fawcett, R., Langhammer, M.: Hyperblock floating point: Generalised quantization scheme for gradient and inference computation. In: 2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). IEEE (2023). https://doi.org/10.1109/wacv56688.2023.00630
13. Netzer, Y., Wang, T., Coates, A., Bissacco, A., Wu, B., Ng, A.Y.: Reading digits in natural images with unsupervised feature learning. In: NIPS Workshop on Deep Learning and Unsupervised Feature Learning 2011 (2011), http://ufldl.stanford.edu/housenumbers/nips2011_housenumbers.pdf
14. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A.C., Fei-Fei, L.: ImageNet Large Scale Visual Recognition Challenge. International Journal of Computer Vision (IJCV) (2015). https://doi.org/10.1007/s11263-015-0816-y
15. Sun, J., Fu, Y., Li, S., He, J., Xu, C., Tan, L.: Sequential human activity recognition based on deep convolutional network and extreme learning machine using wearable sensors. Journal of Sensors (2018). https://doi.org/10.1155/2018/8580959
16. Sun, X., Choi, J., Chen, C.Y., Wang, N., Venkataramani, S., Srinivasan, V., Cui, X., Zhang, W., Gopalakrishnan, K.: Hybrid 8-bit floating point (HFP8) training and inference for deep neural networks. Curran Associates Inc., Red Hook, NY, USA (2019), https://dl.acm.org/doi/10.5555/3454287.3454728
17. Xiao, H., Rasul, K., Vollgraf, R.: Fashion-MNIST: a novel image dataset for benchmarking machine learningalgorithms (2017). https://doi.org/10.48550/ARXIV.1708.07747, https://github.com/zalandoresearch/fashion-mnist
18. Yang, G., Zhang, T., Kirichenko, P., Bai, J., Wilson, A. G., De Sa, C.: Swalp: Stochastic weight averaging in low-precision training (2019). https:/doi.org/10.48550/ARXIV.1904.11943
19. Zhou, S., Wu, Y., Ni, Z., Zhou, X., Wen, H., Zou, Y.: DoReFa-Net: Training low bitwidth convolutional neural networks with low bitwidth gradients (2016). https://doi.org/10.48550/ARXIV.1606.06160

HUMAN-CENTERED COMPUTING–RELIABLE MACHINE LEARNING

Ensembling Machine Learning Models for Malware Detection

Based on the ensemble method for a learning machine and the experience of the authors, *P. Galdames, C. Gutiérrez-Soto, and M. Palomino* propose to ensemble several machine-learning algorithms, evaluating their performance in detecting malware, through static analysis and considering the problems inherent to the malware detection. To accomplish this, the authors use several evaluation metrics.

Design of Feedback for a System to Support Project-Based Learning

K. Sasaki and T. Inoue present six requirements for appropriate feedback in distance project-based learning (PBL). The paper also presents an overview of a system for supporting teachers based on these feedback requirements and refers to own publications.

Orientation-Dependent Chord Length Distribution Functions of Bounded Convex Domains

N. G. Aharonyan, and V. K. Ohanyan state that there is no finite set of directions $V = \{\phi_1, ..., \phi_m\}$, such that the corresponding set of 'orientation-dependent' chord length distributions determines a bounded convex domain uniquely. It is also discussed whether the results given in this article can be used to reconstruct convex bodies from the orientation-dependent chord length distribution function for a finite set of directions.

Assessing Glaucoma Online Tools

N. Baloian and W. Luther present and evaluate various approaches, continuous or simple score-based risk models for estimating the 5-year risk that an individual with ocular hypertension will develop Primary Open Angle Glaucoma (POAG), the leading global cause of irreversible blindness.

Color Image Enhancement with Quaternion Fourier Transform-Based Alpha-Rooting

A. Vardazaryan and A. Grigoryan demonstrate that their quaternion-based approach is more effective at preserving the color relationships and features of an image, offering a significant advancement in the field of image enhancement.

Novel Gradient-Based Retinex Method for Image Enhancement

A. Bayramyan and A. Grigoryan use the gradient-based Retinex method for enhancing drone images and confirm that the applied technique significantly improves the visual quality of both grayscale and color images. A comparative analysis with standard and gradient-based histogram equalization techniques is provided.

Fairness in the Use of Medical Online Tools

W. Luther and A. Harutyunyan focus on the concept of fairness in the use of medical risk assessment tools. The paper discusses the various forms of fairness and the biases that can affect them, such as data quality issues, model inadequacies, algorithmic performance, and stakeholder collaboration. The authors propose a set of guidelines for system requirements to address these fairness concerns, illustrating their application using the example of a tool for estimating the 5-year risk of developing Primary Open Angle Glaucoma (POAG).

Exploring Design Aspects of an AI-supported Farming Platform

A. Mikayelyan and A. Harutyunyan present a prototypical concept for a comprehensive agricultural technology platform that aims to modernize agricultural practices by integrating new approaches of (generative) artificial intelligence. Innovative features allow optimizing crop cultivation, irrigation management and strategic planning for agricultural companies and regulatory actors.

Ensembling Machine Learning Models for Malware Detection

Patricio Galdames[1,2][0000−0003−3051−2413],
Claudio Gutierrez-Soto[3][0000−0002−7704−6141], and
Marco A. Palomino[4][0000−0001−7850−416X]

[1] Universidad San Sebastián, Concepción, Chile, patricio.galdames@uss.cl
[2] Whitecliffe College, Christchurch, New Zealand, 20230676@mywhitecliffe.com
[3] Universidad del Bío-Bío, Concepción, Chile, cogutier@ubiobio.cl
[4] University of Aberdeen, Scotland, UK, marco.palomino@abdn.ac.uk

Abstract. Malware poses serious problems for individuals and businesses worldwide. No matter how much we do to prevent it, attackers continue to find ways to challenge preventive strategies. Thus, we propose a new ensemble learning methodology to enhance malware detection. Our proposal improves predictive accuracy and generalization capabilities by using support vector machines, random forests, and neural networks, all of them trained with publicly available databases and statically extracted features. We have also explored meta-learner construction approaches, such as stacking and weighted voting, to optimally integrate our base detectors, and we have evaluated our ensemble comprehensively, using F1 score, Matthew's correlation coefficient, and informedness metrics, among others. The results demonstrate the potential of ensemble learning for a robust malware defense and highlight the value of informative features, adaptive retraining, and comprehensive evaluation. Clearly, our approach offers better performance than its individual components, particularly in minimizing false positives and negatives.

Keywords: Malware Detection, Ensemble Learning, Machine Learning, Static Analysis, Meta-Learners

1 Introduction

The fourth industrial revolution, marked by the growing integration of sensors into products and processes, has led to a massive increase in data generation. Organizations must deploy advanced technologies to process these data effectively. However, the value of organizational data has also attracted specialized hackers, known as Advanced Persistent Threats (APTs), who are typically a state, or state-sponsored group, and seek to steal data, or gain unauthorized access, for various purposes. Manufacturing companies in the United States are particularly vulnerable, as they have innovative developments that attract APT groups and often lack adequate security controls [16, 45].

Malware, defined as any program or code designed to harm an organization without its knowledge or consent [1], poses significant threats to the confidentiality and availability of data. For example, ransomware attacks, a type of malware, can have devastating financial consequences, as exemplified by the case of MKS Instruments, which suffered losses exceeding $200 million USD [24].

Conventional malware detection relies on the identification of digital signatures or fingerprints within infected files. However, this approach has limitations, especially in the detection of novel malware strains or those employing sophisticated obfuscation techniques. Recent studies [17, 34] have shown the potential of machine learning algorithms to detect unknown malware by analyzing unexpected behavior during runtime. Despite these advances, several challenges remain unsolved in the field of malware detection.

Among the challenges that malware detection must deal with are those that can hinder the development and evaluation of effective techniques, leading to inconsistent or unreliable results. This includes the need for reliable comparison of results across different studies, the availability of unbalanced malware training databases, and issues related to overfitting and generalization of machine learning models. Addressing these obstacles is crucial for progressing the field of malware detection and improving the security of computer systems.

The contributions of this work are threefold:

1. To enable a reliable comparison of results, we use four public databases to train and evaluate our models. Using widely used databases allows for more meaningful comparisons with other related research and validates the effectiveness of the proposed ensemble learning approach.
2. To tackle the problem of unbalanced malware training databases, we employ several evaluation metrics, including Accuracy, Precision Malware, Precision Goodware, Recall Malware, Recall Goodware, F1 Score Malware, F1 Score Goodware, Matthew's correlation coefficient (MCC) and Informedness (INF). These metrics provide a comprehensive assessment of performance.
3. To address overfitting and generalization, we demonstrate that the collaboration of multiple classifiers can be effective. Furthermore, we identified an opportunity for improvement in the development and testing of new voting-based combination techniques.

By integrating several detectors into an ensemble, we compensate for individual shortcomings and leverage complementary strengths. This approach has the potential to improve malware detection accuracy, reduce false positives and negatives, and increase generalization against novel threats, compared to any single, independent detector.

The remainder of this paper is organized as follows: Section 2 explains the problem of malware detection using multiple detectors. Section 3 presents a review of the related literature on malware detection. Section 4 describes the methodology used in this study, including our chosen databases and the proposed collaborative detection approach. Section 5 states our experimental results and a discussion of the findings. Finally, Section 6 concludes and outlines potential future research directions.

2 Problem Statement

This study aims to address the challenge of detecting malware by combining the capabilities of multiple detectors through static analysis. Research has highlighted that machine learning models often struggle with adaptability and generalization when faced with emerging malware strains and novel detection methodologies [1, 30, 34, 50, 55].

Our proposed solution involves adopting a collaborative strategy that integrates various detectors, including those specialized in recognizing older and newer malware strains. This approach aims to enhance the individual performance of these detectors and mitigate the challenges associated with overfitting. However, the incorporation of multiple detectors presents several challenges, including the careful selection of a diverse group of detectors and the choice of an appropriate mechanism to amalgamate their responses.

Although a hybrid solution combining dynamic and static analysis may be optimal, our current focus is exclusively on static analysis—this is due to the absence of a controlled dynamic analysis environment. Static analysis itself faces notable challenges, such as the need for reverse engineering techniques and the meticulous identification of relevant features to achieve accurate detection rates, while minimizing false positives.

3 Literature Review

We will begin by describing basic malware concepts and ensembling methods; then, we will offer an overview of the state-of-the-art on malware detection and machine learning models.

3.1 Background

The following concepts are provided to facilitate the understanding of the literature review.

Portable Executable (PE): This is the standard format used for executables in Microsoft Windows. It contains information such as code, resources, and imports and exports, which are essential for malware analysis [35].

Behavior: This refers to analyzing the actions performed by a program while running—in other words, its behavior—to determine whether it is benign or malign. Actions such as introducing changes to the file system, or attempting to alter the network settings can certainly raise suspicions [47].

Ransomware: A type of malware that encrypts the victim's files and demands a ransom in exchange for the decryption key [19].

Signature: A signature is a way to identify if a given file has a particular type of content. Security researchers often use signatures to identify patterns in a file, and determine if a file is malicious, recognize suspected behavior, and detect a malware family [31].

Hash: A unique value generated by processing a file through a cryptographic hash function, which serves as a digital identifier for the file. It is used to identify and verify known malware through hash databases [21].

Static analysis: Examining a program without executing it by analyzing its code, strings, libraries, etc., which is helpful in extracting signatures and other indicators [26].

Dynamic analysis: Executing a program in a controlled environment to observe its behavior in real-time and look for malicious activities. The controlled environment is meant to be an isolated environment [27].

Ensembling techniques combine multiple models to increase the accuracy and improve the robustness of the predictions. Such techniques can be classified into four main categories:

Bagging: Also known as bootstrap aggregating, combines multiple models—usually decision trees—trained with bootstrap databases to reduce variance and improve generalization [10]. Examples include random forest [11] and bootstrap aggregating for time series (BATS) [7].

Stacking: Also known as stacked generalization, combines the predictions of several base-level models using a higher-level model—or meta-model—to improve prediction accuracy [49].

Voting: An ensemble technique to combine the predictions of multiple models and select the most common, or confidence-weighted, prediction [44].

Boosting: This is a method to train models sequentially. Each model focuses on correcting the errors of the previous model. Examples include AdaBoost [22] and gradient boosting [13].

3.2 Related Work

Solutions for malware detection can be broadly categorized into signature-based and behavior-based approaches. Signature-based approaches involve static analysis of executable files, scrutinizing the program's assembly code to identify suspicious components. Behavior-based approaches analyze programs during runtime in an isolated environment, monitoring resource consumption to distinguish between legitimate and anomalous behavior [18].

Numerous machine learning-based solutions for malware detection have been proposed [14, 23, 28, 36, 43], using algorithms such as support vector machines (SVMs) [36], decision trees (DT) [28], and random forests [23]. Schultz et al. [43] pioneered the application of data mining methods for malware detection, incorporating static features such as the header, string, and byte sequence of the PE. Cohen et al. [14] combined a rule-based algorithm with a naïve Bayesian algorithm to identify pertinent features and patterns within byte sequences. Kolter et al. [28] introduced the N-gram technique as an alternative to sequence features, using decision trees. Saxe et al. [42] used a neural network for malware detection, training it with features such as an entropy histogram, contextual data calls, metadata, and DLL imports.

Nataraj et al. [37] converted the binary data of malware into grayscale images and used k-nearest neighbors (KNN) with Euclidean distance for classification. Other neural network architectures, such as convolutional networks (CNN) [40], graph convolutional networks (GCN) [38], deep-belief networks (DBN) [54], long short-term memory (LSTM) [56], VGG16 [25], and generative adversarial networks (GAN) [32], have also been employed for malware detection. However, it remains challenging to ensure the generalizability of these models [55].

Ensemble methods that integrate multiple learning algorithms have been extensively studied to improve prediction accuracy and mitigate overfitting [17]. In the bagging category, researchers have employed random forests and extremely randomized trees [29, 57, 52]. Damaševičius et al. [15] employed ensemble learning methods—including random forest, which is a typical bagging method—to combine classifiers based on neural networks for Windows PE malware detection.

In the boosting category, Al Sarah et al. [3] have evaluated ensemble methods, such as the gradient boosting decision tree, light GBM, and XGBoost. Feng et al. [20] found that stacking achieves the best classification performance in the detection of Android malware, while Yan et al. [51] combined CNN and LSTM networks for the detection of Android malware. Taha and Barukab [46] investigated the classification of Android malware using a random forest classifier as a meta-learner—the parameters were optimized using a genetic algorithm. Roy et al. [41] proposed a hybrid stacked ensemble learning framework with feature engineering schemes for the analysis of obfuscated malware. Rana and Sung [39] evaluated ensemble learning for Android malware detection.

As far as voting is concerned, Wang et al. [48] used an ensemble of SVM, KNN, naïve Bayes, CART, and random forest, using majority voting. Yerima et al. [53] proposed DroidFusion, a framework that generates a model by training base classifiers at a lower level and then applies ranking-based algorithms on their predictive accuracies at a higher level.

4 Methods

Herein, we explain the methods selected to address the problem statement. We specify which machine learning models will be trained, what ensemble techniques will be used, and what evaluation metrics we plan to utilize.

4.1 Data Collection, Pre-processing, and Model Training

To achieve our objective, we followed the next five steps:

Database selection: We have identified four public databases to carry out our work. These databases have been used in recent studies [2, 5, 39, 12, 53]. By training and testing the models on multiple and diverse databases, the study assesses performance and generalizability across various malware types and data sources, reducing the risk of overfitting and providing a robust evaluation of the models' reliability. The databases are:

- *DEBRIN* [8], which contains 5,560 applications from 179 different malware families and 123,453 benign applications, with 215 observable string features extracted from the samples.
- *AndroZoo* [4], which comprises 267,342 benign and 80,102 malicious applications from an original database of more than 10 million applications—AndroZoo maintains the same distribution as the original database.
- *UCI* [6], which covers 373 samples, of which 301 are malicious and 72 are safe, as described by 531 unlabelled features.
- *Benign and Malware PE database* [33], which includes 19,612 samples, where 14,600 are malware and 5,010 are goodware—that is, benign, safe software. Each sample is described by 79 PE features.

Pre-processing: This phase evaluates the compatibility of the chosen database to train and test machine learning models. It involves managing missing data, balancing unbalanced classes, and selecting features.

Model selection and training: This requires identifying the individual, most-cited models for malware detection to function as benchmark references. Based on the literature review, the following algorithms have been selected for testing: SVM, naïve Bayes, decision tree, KNN, random forest, sequential neural network, AdaBoost, extra trees, gradient boosting, voting classifier, XGBoost, LightGBM, CatBoost, and bagging classifier. These algorithms were chosen due to their prevalence in malware detection research and their various classification approaches, providing a comprehensive set of benchmarks for comparison with ensemble methods.

Ensemble machine learning selection: The strategy encompasses studying the ensemble techniques proposed in the literature and selecting a few to combine the best performing individual malware detector. The chosen base models were combined to build a meta-model using the following algorithms: simple voting, weighted voting based on predictor accuracy rate, and weighted voting calculated with logistic regression.

Parameter refinement and model testing: Creating ensemble models comprises meticulous parameter tuning to optimize performance. We employed cross-validation at this stage.

4.2 Evaluation Metrics and Validation Strategy

Our evaluation approach involved an analysis based on several metrics, namely, accuracy, precision, F1 score, recall, MCC, and INF. With all these metrics, we expected to gather a comprehensive understanding of performance. The formulae stated below elucidate the calculation of each metric involved in our study [9]:

- $Accuracy = \dfrac{TP + TN}{TP + TN + FP + FN}$
- $Precision = \dfrac{TP}{TP + FP}$
- $Recall = \dfrac{TP}{TP + FN}$

- $F1 = \dfrac{2 * Precision * Recall}{Precision + Recall} = \dfrac{2 * TP}{2 * TP + FP + FN}$
- $MCC = \dfrac{TP * TN - FP * FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$
- $INF = Sensitivity + Specificity - 1 = \dfrac{TP}{TP + FN} + \dfrac{TN}{TN + FP} - 1$

Ranking algorithms have been developed to sort the base classifiers and determine which will provide the final response. These include the rank according to the balance of precision and class difference (BPCD), the rank according to F1, and the rank according to MCC and INF.

The models were validated using a split training and testing set, where the training set consisted of 70% of the data and the remaining 30% was left to the test set. Cross-validation was employed to ensure robustness by minimizing potential bias. Using a comprehensive evaluation approach allowed us to have a holistic assessment of the performance of the models.

5 Results

This section explores the experimental setup. Real-world Android and Windows PE malware datasets were used to train individual and ensemble classifiers, combining the algorithms from the previous section with stacking and voting to create an optimized detector.

The simulations were run on an ASUS laptop with a 12th generation Intel Core i7-12700H processor at 2.3GHz, 32GB of DDR4 RAM, and a 64-bit Windows 11 Home operating system.

Python 3.11.5 was used for the development and execution of the experiments, along with core libraries like NumPy, SciPy, Matplotlib, and Pandas. The development environment consisted of IPython 8.15.0, Ipykernel 6.25.0, Jupyter Notebook 6.5.4, and JupyterLab 3.6.3, which facilitated easy editing and execution of Python notebooks.

To maintain integrity and robustness, we focus on the more comprehensive DEBRIN, AndroZoo, and Benign and Malware PE databases. UCI was initially considered for the study but ultimately excluded, due to its small size, which is insufficient for reliable training and evaluation—it may lead to overfitting and poor generalization to real-world scenarios.

5.1 Top Individual Models

Table 1 presents the top individual models across Debrin, AndroZoo, and PE databases. The random forest classifier consistently achieves high accuracy, precision, recall, and F1 scores for malware detection. The decision tree classifier also performs well, especially in terms of the recall of malware samples.

Model	Accuracy	Precision Malware	Recall Malware	F1 Malware
Random Forest (Debrin)	0.989614	0.964620	0.798450	0.873704
Random Forest (AndroZoo)	0.921466	0.900468	0.767162	0.828487
Random Forest (PE)	0.989804	0.989559	0.996964	0.993239
Decision Tree (Debrin)	0.982754	0.797342	0.826870	0.811839
Decision Tree (AndroZoo)	0.879586	0.750052	0.769368	0.759587
Decision Tree (PE)	0.983431	0.989803	0.988124	0.988962

Table 1. Top individual model metrics across datasets

Table 2 shows the average performance of the base models. The random forest classifier achieves the highest accuracy, precision, and F1 scores for the detection of malware and goodware. The decision tree classifier also performs well, with the best average recall for malware detection. However, most other base models struggle to detect malware effectively, as evidenced by their low precision, recall, and F1 scores for the malware class.

Model	Accuracy	Precision Malware	Precision Goodware	Recall Malware	Recall Goodware	F1 Malware	F1 Goodware	MCC	INF
Random Forest	0.969961	0.958209	0.971972	0.854192	0.972162	0.916988	0.969677	0.874528	0.87740
Decision Tree	0.948224	0.848066	0.956751	0.860521	0.957883	0.850129	0.958539	0.774029	0.774200
SGD	0.831203	0.583647	0.837690	0.307392	0.981001	0.299166	0.814207	0.191799	0.241341
Gaussian NB	0.670508	0.598757	0.634636	0.122554	0.981803	0.194718	0.695974	0.164928	0.184890
SVC	0.765456	0.813700	0.861373	0.712232	0.997124	0.597984	0.886741	0.256058	0.383373
MLP	0.859367	0.743993	0.843407	0.320366	0.994337	0.418469	0.865237	0.370828	0.444137
Multinomial NB	0.790361	0.375732	0.867356	0.499761	0.694966	0.430559	0.668273	0.056223	0.128578
Logistic regression	0.668435	0.228981	0.720006	0.015440	0.990575	0.026937	0.757011	0.035849	0.088189

Table 2. Base models table (average values)

5.2 Ensemble Results

Table 3 shows the performance of the ensemble models. The stacking classifier consistently achieves the highest accuracy and strong precision, recall, and F1 scores for malware detection. The voting classifier also performs well, especially in terms of recall of malware samples in the PE database. Ensemble models demonstrate a better balance between malware and goodware detection than the individual models, as evidenced by the MCC and INF scores.

Model	Accuracy	Precision Malware	Recall Malware	F1 Malware	MCC	INF
Stacking (Debrin)	0.990311	0.954138	0.824289	0.884473	0.882009	0.884167
Stacking (AndroZoo)	0.921589	0.885657	0.784093	0.831786	0.783374	0.784079
Stacking (PE)	0.989549	0.990216	0.995663	0.993064	0.971942	0.971959
Voting (Debrin)	0.960005	0.992366	0.111972	0.201238	0.326469	0.532069
Voting (AndroZoo)	0.780825	0.887370	0.130040	0.226840	0.287703	0.394409
Voting (PE)	0.990059	0.989892	0.996964	0.993407	0.973312	0.973338

Table 3. Ensemble model metrics across datasets

Table 4 presents the average performance of the ensemble models. The stacking classifier achieves the highest accuracy, precision, recall, and F1 scores for malware and goodware detection. The extra trees, XGBoost, and CatBoost classifiers also perform well, with high accuracy and F1 scores. The ensemble models generally maintain strong malware detection capabilities, unlike the base models. The stacking classifier demonstrates the most balanced performance, as evidenced by its high MCC and INF scores.

Model	Accuracy	Precision Malware	Precision Goodware	Recall Malware	Recall Goodware	F1 Malware	F1 Goodware	MCC	INF
AdaBoost	0.922235	0.805845	0.926577	0.620425	0.965890	0.665690	0.944795	0.631549	0.635936
ExtraTrees	0.967011	0.956806	0.967848	0.833272	0.969199	0.887031	0.968283	0.838969	0.839751
Gradient Boosting	0.931184	0.902946	0.936004	0.625054	0.976423	0.726933	0.957025	0.687589	0.694193
Voting	0.882295	0.893019	0.887678	0.383659	0.984940	0.425370	0.924307	0.505891	0.625226
XGB	0.961417	0.936941	0.963909	0.839536	0.969934	0.879058	0.963047	0.855707	0.857088
LGBM	0.956564	0.927029	0.961072	0.713984	0.975475	0.800860	0.952469	0.751271	0.754238
CatBoost	0.965072	0.941162	0.966105	0.822152	0.969493	0.872650	0.963334	0.851559	0.852983
Bagging	0.960097	0.936585	0.964003	0.796211	0.963996	0.860636	0.961599	0.815256	0.816340
Stacking	0.979166	0.962070	0.980363	0.862876	0.972552	0.905873	0.977447	0.882282	0.883662

Table 4. Ensemble models table (average values)

The significant difference in the recall malware values between the base models—Table 2—and the voting ensemble—Table 3—can be attributed to the nature of the voting ensemble and the characteristics of the base models.

Most base models, except for Random Forest and Decision Tree, have relatively low recall malware values, indicating that they struggle to correctly identify a large portion of malware samples. The voting ensemble, which combines the predictions of multiple base models, is more conservative in classifying samples as malware, potentially requiring a higher consensus among the base models. As a result, the voting ensemble may miss more malware samples (lower recall) to maintain higher precision and avoid false positives. In contrast, the stacking ensemble (Table 3) achieves a better balance between recall and precision for malware detection, effectively combining the predictions of the base models and improving overall performance.

5.3 Dataset Variances

The performance of the ensemble models varies among the different databases. For Debrin, the stacking classifier achieves an MCC of 0.882009, indicating a strong balance between malware and goodware detection. For AndroZoo, the stacking classifier achieves an MCC of 0.783374, which is lower than in the case of Debrin, but still demonstrates good performance. For the PE database, the highest MCC of 0.973312 is achieved by the extra trees classifier, closely followed by the stacking classifier with an MCC of 0.971942.

5.4 Discussion

In malware databases, random forest consistently emerged as the best individual model, with accuracy, precision, and recall above 95% in most cases. However, all base learners showed bias, and struggled to detect the minority positive malware class effectively.

Ensemble models improved individual results, both in terms of overall accuracy and the mitigation of class imbalance. The stacking classifier ensemble achieved the best overall performance, leveraging multiple base models to achieve a precision greater than 99% in some cases. Largely, the gap between precision and recall in malware and goodware was reduced to within 5-10%.

Other ensemble techniques such as voting, XGBoost, LightGBM, and CatBoost also showed promising results. Combining models via ensembling helped to overcome limitations like severe data skews, which impacted individual classifiers. Both the malware detection rate and false positives were improved by model aggregation.

6 Conclusions

We have explored ensemble learning for malware detection by identifying newly published malware databases, implementing prominent detection techniques, and testing collaborations of multiple models. Comparative analysis revealed an improved performance of integrated learners over individual techniques.

The study demonstrated ensemble learning's potential for more robust malware defense. Nevertheless, success depends on factors like representative databases, informative features, adaptive retraining, and multimetric evaluations. With deliberate design, integrated systems provide a promising way to match adversaries.

Opportunities remain to evaluate non-linear voting combinations and incorporate sample-specific information, such as probability vectors from multiclass neural network classifiers. Potentially, this could improve confidence when classifying new samples. Undoubtedly, integrating dynamic weighting in an optimal manner merits future work.

Specific findings include consistently high performance of the random forest classifier and stacking to assemble heterogeneous models. However, exploration is still nascent. Solidifying the capabilities of combinations of detectors remains a task for future work, as attacks keep growing and becoming more potent.

7 Acknowledgments

This work was made possible with the support of Universidad San Sebastián, Concepción, Chile. The authors express their sincere gratitude to Professor Joseph Dang of Whitecliffe College, New Zealand, for his valuable feedback and constructive comments. Marco Palomino acknowledges the funding provided by the University of Aberdeen to support his work on this publication.

References

1. Akhtar, M., Tao, F.: Malware analysis and detection using machine learning algorithms. Symmetry **14**(11), 2304 (2022)
2. Akhtar, M., Feng, T.: Evaluation of machine learning algorithms for malware detection. Sensors **4**(2) (2023)
3. Al Sarah, N., Rifat, F.Y., Hossain, M.S., Narman, H.S.: An efficient android malware prediction using ensemble machine learning algorithms. Procedia Computer Science **191**, 184–191 (2021)
4. Allix, K., Bissyandé, T., Klein, J., Le Traon, Y.: Androzoo: Collecting millions of android apps for the research community. In: Proceedings of the 13th International Conference on Mining Software Repositories. pp. 468–471. MSR '16 (2016)
5. Alqahtani, A., Azzony, S., Alsharafi, L., Alaseri, M.: Web-based malware detection system using convolutional neural network. Digital **3**(3), 273–285 (2023)
6. AnastaRumao, P.: Using two-dimensional hybrid feature dataset to detect malicious executables. International Journal of Innovative Research in Computer and Communication Engineering **4**(7) (2016)
7. Andrade, D., Galeano, P., Mendoza, C.: Bats: A Bayesian time series model for classification and regression. Machine Learning **106**(10), 1595–1617 (2017)
8. Arp, D., Spreitzenbarth, M., Hübner, M., Gascon, H., Rieck, K.: Drebin: Effective and explainable detection of Android malware in your pocket. In: Proceedings of the 21st Network and Distributed System Security Symposium (NDSS). pp. 23–26. Internet Society, San Diego, CA, USA (2014). https://doi.org/10.14722/ndss.2014.23247

9. Banerjee, R.: Understanding Accuracy, Recall, Precision, F1 Scores, and Confusion Matrices (2021), https://towardsdatascience.com/understanding-accuracy-recall-precision-f1-scores-and-confusion-matrices-561e0f5e328c [Accessed: 14 June 2024]
10. Breiman, L.: Bagging predictors. Machine learning **24**(2), 123–140 (1996)
11. Breiman, L.: Random forests. Machine learning **45**(1), 5–32 (2001)
12. Ceschin, F., Botacin, M., Murilo-Gomes, H., Pinagé, F., Oliveira, L S.., Grégio, A.: Fast & furious: On the modelling of malware detection as an evolving data stream. Expert Systems with Applications p. 118590 (2023)
13. Chen, T., Guestrin, C.: XGBoost: A scalable and accurate implementation of gradient boosting machines. Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining pp. 785–794 (2016)
14. Cohen, W.: Fast effective rule induction. In: Prieditis, A., Russell, S. (eds.) Machine Learning Proceedings, pp. 115–123. Morgan Kaufmann, San Francisco (CA) (1995)
15. Damaševičius, R., Venčkauskas, A., Toldinas, J., & Grigaliūnas, Š.: Ensemble-based classification using neural networks and machine learning models for windows PE malware detection. Electronics 10(4), 485 (2021). https://doi.org/10.3390/electronics10040485
16. Death, D.: Information Security Handbook: Develop a threat model and incident response strategy to build a strong information security framework. Packt, New York NY (2017)
17. Dong, X., Yu, Z. and Cao, W., Shi, Y., Ma, Q.: A survey on ensemble learning. Front. Comput. Sci. **14**, 241–258 (2020)
18. Egele, M., Scholte, T., Kirda, E., Kruegel, C.: A survey on automated dynamic malware-analysis techniques and tools. ACM Computing Surveys **44**(2), 1–42 (2008)
19. FBI: Ransomware (2023), https://www.fbi.gov/how-we-can-help-you/safety-resources/scams- and-safety/common-scams-and-crimes/ransomware
20. Feng, P., Ma, J., Sun, C., Xu, X., Ma, Y.A.: Novel dynamic android malware detection system with ensemble learning. IEEE Access **6**, 30996–31011 (2018)
21. Frankenfield, J.: Cryptographic hash functions: Definition and examples, updated 2024, https://www.investopedia.com/news/cryptographic-hash-functions/
22. Freund, Y., Schapire, R.: Experiments with a new boosting algorithm. Machine learning: Proceedings of the 13th International Conference pp. 148–156 (1996)
23. Hassen, M., Carvalho, M., Chan, P.: Malware classification using static analysis based features. In: 2017 IEEE Symposium Series on Computational Intelligence (SSCI). pp. 1–7 (2017)
24. Helyome, K.: Ransomware attacks in manufacturing and what business leaders fear most (2023), https://blogs.blackberry.com/en/2023/03/ransomware-attacks-in-manufacturing-and-what-business-leaders-fear-most
25. Huang, X., Ma, L., Yang, W., Zhong, Y.: A method for windows malware detection based on deep learning. J Sign Process Syst **93**, 265–273 (2020)
26. InfoSec Institute, Inc.: Resource Center — Malware Analysis — Static Analysis–Part 2 (2017), https://resources.infosecinstitute.com/topics/malware-analysis/malware-behavioral-code-analysis-part-2/ [Accessed on 14 June 2024]
27. Jones, S.: What is dynamic malware analysis? (2023), https://www.bitdefender.com/blog/businessinsights/what-is-dynamic-malware- analysis/
28. Kolter, J., Maloof, M.: Learning to detect and classify malicious executables in the wild. Journal of Machine Learning Research **7**, 2721–2744 (December 2006)

29. Kumar, R., Subbiah, G.: Explainable machine learning for malware detection using ensemble bagging algorithms. In: Proceedings of the 14th International Conference on Contemporary Computing (IC3). pp. 453–460 (August 2022)
30. LinkedIn Community: What are the main challenges and limitations of machine Learning for malware detection? (2023), https://www.linkedin.com/advice/1/what-main-challenges-limitations-machine-learning [Accessed: 14 June 2024]
31. MalwareBytes—Cyberprotection for every one: Signature (2023), https://www.malwarebytes.com/glossary/signature [Accessed: 14 June 2024]
32. Martins, N., Cruz, J., Cruz, T., Henriques, P.: Adversarial machine learning applied to intrusion and malware scenarios: A systematic review. IEEE Access **8**, 35403– 35419 (2020)
33. Mauricio, A.: Benign and malicious PE Files Dataset for malware detection (2018), https://www.kaggle.com/datasets/amauricio/pe-files-malwares [Accessed: 14 June 2024]
34. Menzli, A.: Building trust in machine learning malware detectors (2020), https://towardsdatascience.com/building-trust-in-machine-learning-malware- detectors-d01f3b8592fc [Accessed: 14 June 2024]
35. Microsoft: PE Format (2023), https://learn.microsoft.com/en-VT/XJOEPXT/XJO32/debug/pe-format [Accessed on 14 June 2024]
36. Naeem, H., Guo, B., Naeem, M.: A light-weight malware static visual analysis for IOT infrastructure. In: 2018 International Conference on Artificial Intelligence and Big Data (ICAIBD). pp. 240–244 (2018)
37. Nataraj, L., Karthikeyan, S., Jacob, G., Manjunath, B.S.: Malware images: Visualization and automatic classification. In: Proceedings of the 8th International Symposium on Visualization for Cyber Security (2011)
38. Pei, X., X., Yu, L., Tian, S.: Amalnet: A deep learning framework based on graph convolutional networks for malware detection. Computers & Security **93**, 101792 (2020)
39. Rana, S., Sung, A.H.: Evaluation of advanced ensemble learning techniques for android malware detection. Vietnam Journal of Computer Science **7**(2), 145–155 (2020)
40. Ren, Z., Wu, H., Ning, Q., Hussain, I., Chen, B.: End-to-end malware detection for android IOT devices using deep learning. Ad Hoc Networks **101**, 102098 (2020)
41. Roy, K.S., et al.: MalHyStack: A hybrid stacked ensemble learning framework with feature engineering schemes for obfuscated malware analysis. Elsevier Intelligent Systems with Applications **20**(11), 200283 (2023)
42. Saxe, J., Berlin, K.: Deep neural network based malware detection using two dimensional binary program features. In: 2015 10th International Conference on Malicious and Unwanted Software (MALWARE). pp. 11–20 (2015)
43. Schultz, M., Eskin, E., Zadok, F., Stolfo, S.: Data mining methods for detection of new malicious executables. In: Proceedings 2001 IEEE Symposium on Security and Privacy. S&P 2001. pp. 38–49 (2001)
44. scikit-learn Developers: Ensembles: Gradient boosting, random forests, bagging, voting, stacking (2023), https://scikit-learn.org/stable/modules/ensemble.html [Accessed on14 June 2024]
45. Spadafora, A.: Industry 4.0 suffering major security issues (2019) https://www.techradar.com/news/industry-40-suffering-major-security-issues
46. Taha, A., Barukab, O.: Android malware classification using optimized ensemble learning based on genetic algorithms. Sustainability **14**(21), 14406 (2022). https://doi.org/10.3390/su142114406

47. Virgilito, D.: Resource Center—Malware Analysis — Common Malware Behavior (2019), https://resources.infosecinstitute.com/topics/malware-analysis/common-malware-behavior/ [Accessed: 14 June 2024]
48. Wang, W., Y., L., Wang, X., Liu, J., Zhang, X.: Detecting android malicious apps and categorizing benign apps with ensemble of classifiers. Futur. Gener. Comput. Syst **78**, 987–994 (2018)
49. Wolpert, D.: Stacked generalization. Neural Networks **5**(2), 241–259 (1992)
50. Xiao, M., Wu, Y., Zuo, G., Fan, S., Yu, H., Shaikh, Z., Wen, Z.: Addressing over- fitting problem in deep learning-based solutions for next generation data-driven networks. Wireless Communications and Mobile Computing **2021** (2021)
51. Yan, J., Qi, Y., Rao, Q.: Detecting malware with an ensemble method based on deep neural networks. Secur. Commun. Netw. **2018**, 1–16 (2018)
52. Ye, Y., Chen, L., Wang, D.L.T., Jiang, Q., Zhao, M.: SBMDS An interpretable string based malware detection system using SVM ensemble with bagging. Journal in Computer Virology **5**, 283–293 (2009)
53. Yerima, S., Sezer, S.: DroidFusion: A Novel Multilevel Classifier Fusion Approach for Android Malware Detection. IEEE Transactions on Cybernetics **49**(2), 453–466 (2018)
54. Yuxin, D., Siyi, Z.: Malware detection based on deep learning algorithm. Neural Comput & Applic **31**, 461–472 (2019)
55. Zador, A.: A critique of pure learning and what artificial neural networks can learn from animal brains. Nature communications **10**(1), 3770 (2019)
56. Čeponis, D.; Goranin, N. Investigation of Dual-Flow Deep Learning Models LSTM-FCN and GRU-FCN Efficiency against Single-Flow CNN Models for the Host-Based Intrusion and Malware Detection Task on Univariate Times Series. Applied Sciences **10**(7) (2020)
57. Şahin D., Akleylek, S., Kiliç, E.: Linregdroid: Detection of android malware using multiple linear regression models-based classifiers. IEEE Access **10**, 14246–14259 (2022)

Design of Feedback for a System to Support Distance Project-Based Learning

Kosuke Sasaki[1,2][0000−0002−1011−8884] and Tomoo Inoue[3][0000−0003−3600−214X]

[1] Graduate School of Library, Information and Media Studies, University of Tsukuba, Ibaraki, Japan
[2] Faculty of Global Management, Chuo University, Tokyo, Japan
ksasaki@slis.tsukuba.ac.jp
[3] Institute of Library, Information and Media Science, University of Tsukuba, Ibaraki, Japan
inoue@slis.tsukuba.ac.jp

Abstract. This study focuses on project-based learning in a distance environment (distance PBL) and investigates the requirements for appropriate feedback that teachers provide to learners and the support system for teachers' feedback. In distance PBL, it is difficult for teachers to grasp the progress of individual learners, making it difficult for them to provide appropriate feedback, which is necessary for learners to proceed smoothly with their learning activities. Although previous studies have examined who should be given feedback, there has been a lack of consideration regarding what kind of feedback should be given to learners. The purpose of this study is to obtain insights into the appropriate feedback that teachers should give to learners in distance PBL by surveying previous studies. Based on these findings, this study presents six feedback requirements. This paper also presents an overview of a system to satisfy feedback requirements and to support teachers' feedback in distance PBL.

Keywords: Feedback · Assessment · Project-Based Learning · Distance Learning · Self-Regulation · Activity Report

1 Introduction

Distance learning has become increasingly popular. Distance learning has several advantages, including the possibility of taking classes from home without having to attend in-person school [15] and the possibility of learning at the learner's own pace using on-demand videos of classes [1, 17]. This study focuses on project-based learning (PBL) in a distant environment (distance PBL), in which research activities are the primary learning activities in higher education. PBL is a learner-centered learning method that integrates knowledge acquisition with activities aimed at solving real-world problems.

In PBL, it is difficult for teachers to devote sufficient time to each learner, as they generally teach multiple learners. Furthermore, it has been reported that

communication opportunities can be reduced in distant environment compared to face-to-face environment [24, 35]. Thus, in distance PBL, it is difficult for teachers to monitor the progress of all learners individually.

Even if it is difficult to understand the learner's situation, teachers need to provide appropriate feedback to the learner to promote PBL [3]. Thus, this study aims to establish a mechanism that allows teachers to easily grasp the status of individual learners and also allows for easy feedback.

For teachers to grasp learners' learning status, Sasaki et al. examined a method to estimate engagement, a mental indicator of positive attitudes toward learning, by using learners' daily activity reports. Engagement indicates whether a learner ' s learning is progressing well [45]. That is, activity reports provide clues about the status of individual learners. Using this finding, it is easy to identify learners whose engagement is declining and whose learning is presumed to not progress well.

On the other hand, there is a lack of investigation into the appropriate feedback that teachers should provide to learners to carry out smooth PBL. The purpose of this study is to organize the insights of previous studies on feedback to determine what appropriate feedback should be in distance PBL. We also aim to construct a system to support teachers in providing feedback in distance PBL using the findings from our survey.

Based on the findings of this study, this paper presents six requirements for appropriate feedback in distance PBL. This paper also presents an overview of a system for supporting teachers, based on the feedback requirements.

2 What Is Feedback?

In this study, feedback refers to information that facilitates learning for learners in distance PBL. We organized the findings of previous research on what constitutes appropriate feedback in distance PBL, which this study aims to support.

Feedback is used to change the gap between standards and current learning status [8, 41]. Sadler, who summarized formative assessment in learning, argues that feedback contains information about what learners can do and how well they can do it, and allows learners to know what skills they need to learn [44]. Hattie and Timperley, who conducted a comprehensive survey of feedback on learning activities, also argued that providing feedback can enhance learning activities when there is a discrepancy between where the learner is currently headed and where they should be heading in relation to their learning goal [22]. Another study points out that feedback can be used to help learners understand the answers to learning questions [7, 14, 20, 29, 33, 34, 37]. These studies claimed that feedback could close the gap with learning standards.

Feedback in learning is important for learners' acquisition of knowledge and skills [49]. Therefore, feedback is an essential element of instruction [7]. However, as previous studies have pointed out, feedback can have a variety of positive and negative effects in various situations [7, 22, 27, 49]. We first summarized the

characteristics of feedback that have been argued in previous studies to examine how feedback provided by teachers in distance PBL can be supported.

2.1 Feedback Levels

Hattie and Timperley argue that feedback works on four levels: feedback about the task, the process, self-regulation, and the self as a person [22]. Based on this, we summarized the four levels at which feedback works.

Feedback About the Task Feedback about the task is the most common type of feedback provided to ensure that the correct knowledge and skills are acquired. Having the correct information allows the process and self-regulation to occur effectively [22]. However, correct knowledge and skills often vary, depending on the learning context. In particular, in PBL, which is the focus of this study, each learner has a different learning project and task, which requires a different task for each learner and the knowledge and skills required for each task. This is one of the factors that make feedback difficult in PBL because teachers need to check whether these knowledge and skills have been acquired correctly.

It has also been pointed out that feedback that merely provides learners with knowledge and skills may result in poor learning. Kluger and DeNisi, who studied the intervention effects of feedback in learning, pointed out that intervention through excessive feedback may force learners to recognize that they have incorrect knowledge or skills about a task and reduce their cognitive effort in performing the task [27]. Especially in PBL, where learners are expected to carry out learning themselves [32], reducing motivation for the task through feedback should be avoided.

Feedback About the Process of the Task Feedback about the process is related not only to the task process but also to the learning environment [22]. Feedback working at the process level includes, for example, checking whether the goals of learning are not different from those required by the teacher, or examining whether the learning plan has broken down. It has been argued that there is an interaction effect between feedback about the task and the process, which has been reported to be related to task performance and learner confidence [18, 22]. Feedback that acts at both task and process levels is required.

Feedback About Self-Regulation Self-regulation feedback focuses on how learners monitor and regulate their learning according to their learning goals. Hattie and Timperley noted that the following six aspects particularly need attention when providing self-regulation feedback [22].

The first is the learner's ability to provide internal feedback and self-assessment. Learners self-assess in two aspects: self-appraisal, in which they review their abilities and state of knowledge, and self-management, in which they review their learning plans and correct mistakes [40]. In addition, the experience of self-assessment allows learners to perform multiple aspects of their performance,

including the ability to make comparisons relative to the goals and performance of others [22]. Teachers are expected to provide feedback that helps learners develop their self-evaluation skills.

The second is the willingness to expend effort seeking and processing feedback information. Feedback for learners is intended to help them learn more successfully. Reaping the benefits of feedback comes at the cost of the learner's effort to receive feedback, accept the fact that others have evaluated them, and correctly interpret the feedback given to them. If the cost of receiving feedback becomes prohibitive relative to the benefits of the feedback received, learners may stop seeking feedback [22]. Teachers should also consider the costs of receiving feedback.

The third is the learner's level of confidence or certainty in responding to learning questions. It has been noted that when learners are confident in their answers and their answers are correct, they pay little attention to the feedback [30]. Feedback also tends to be ignored if learners are not confident in their answers [29]. As feedback is information that helps close the gap with the criterion, the greatest effect on learners is expected when they are given feedback that they are wrong on answers that they were confident were correct. It has been noted that teachers can achieve effective feedback by providing additional instruction and information, especially for learners who are not confident in their answers [22].

The fourth factor is learners' self-efficacy. Self-efficacy refers to the degree to which one perceives one's perception of their abilities [6]. Kluger and DeNisi argued that feedback increases learners' self-efficacy and enables them to self-regulate more effectively [27]. Self-efficacy and learning performance are related. Learners who receive feedback and increase their self-efficacy can enhance their learning performance through effort [13, 48]. Nicol and Macfarlane-Dick found that feedback is used to increase learners' self-efficacy and argued that it can be used to result in significant benefits to the learner's knowledge and skills, and ultimately to their level of education [39].

The fifth factor is the learner's attribution of success or failure. It has been suggested that the failure to link feedback to factors of success or failure in learning may have a negative impact on self-efficacy and task performance [22]. For example, if feedback is unclear, self-handicapping may occur because learners are unable to identify the factors of their success or failure, creating an uncertain self-image of their future [50]. However, Hattie and Timperley cautioned against attributing factors of success or failure to a learner's effort or ability. For example, feedback that focuses on learners ' efforts is expected to be effective in the early stages of learning. On the other hand, in the later stages of learning, the effort involved is relatively low, so it is more reliable to give feedback that focuses on the learner's ability, noting that learning progress is consistent with the learning achievement goal [22].

The sixth is proficiency in asking for help. When learners ask teachers for help through feedback, they request two main types of help: executive help seeking and instrumental help seeking. In executive help seeking, learners seek answers

directly to save time and effort in task completion, which is associated with task-level and process-level feedback. On the other hand, in instrumental help seeking, the learner does not seek a direct solution but rather looks to the teacher for hints to advance learning, leading to self-regulation-level feedback. However, many students tend to be reluctant to ask teachers because of their low self-efficacy, fear of threatening their self-esteem, and embarrassment [22, 26, 38, 42, 43]. To promote help-seeking behavior, measures that lower mental hurdles to the desire for help are important.

It has been noted that learners who are ineffective in learning do not have many self-regulation strategies [22]. Feedback about self-regulation has also been shown to have a stronger impact on learning than any other level of feedback [14]. Feedback at the self-regulation level is one of the key elements in PBL, which is the focus of this study as it allows learners to regulate and promote their learning activities.

Feedback about the Self as a Person Self-level feedback, such as "good boy/girl" or "great effort," is often used in place of task, process, and self-regulation level feedback [22]. Although self-level feedback is primarily used to express positive evaluations and feelings [12], it does not contain task-relevant information, and has been reported to have limited effects on learning [28]. It has also been noted that self-level feedback may influence learner motivation [9, 22].

On the other hand, praise for the task or process, rather than just praise for the learners, may lead to an increase in their self-efficacy. For example, praise that mentions the completion of the task or how the task was completed, such as "You did a great job completing the task using this method," is expected to have a positive effect on performance [22].

Self-level feedback does not directly affect learning, but may indirectly guide learners to facilitate learning by working on their mental parts.

2.2 Feedback Strategy

We then summarized the arguments of previous research on feedback strategies. First, we reviewed the principles of feedback. Next, we organized the findings of previous studies on each of the following points: (1) what information should be included in feedback, (2) when feedback should be provided, and (3) who should be provided feedback.

Feedback Principles Nicol and Macfarlane-Dick derived the following principles of good feedback based on their literature review [39].

- Helps clarify what good performance is (goals, criteria, expected standards).
- Facilitates the development of self-assessment (reflection) in learning.
- Delivers high-quality information to students about their learning.
- Encourages teacher and peer dialogue around learning.

- Encourages positive motivational beliefs and self-esteem.
- Provides opportunities to close the gap between current and desired performance.
- Provides information to teachers that can be used to help shape the teaching.

It has also been noted that continuous feedback during the learning period is more powerful than one-time feedback and that it is important to create a feedback loop [23, 39, 11]. For effective feedback, Brooks et al. proposed the following feedback principles [11].

- Clarify expectations and standards for the learner.
- Schedule ongoing, targeted feedback within the learning period.
- Foster practices to develop self-regulation.
- Provide feedforward opportunities to implement the feedback and close the feedback loop.

Common to both these principles is the need for teachers to clarify the standards of learning for learners. Presenting goals and standards to learners is an important prerequisite for effective feedback [10, 22, 39, 11]. The sharing of learning intentions through feedback between learners and teachers can be used to help learners understand where they stand in their learning and to close the gap with their goals [51].

The feedback principles of Brooks et al. also include the concepts of feedforward as well as feedback. As Hattie and Timperley, and Boud and Molloy also argue, feedforward is information that allows learners to take voluntary learning actions and can facilitate the improvement of learning activities [10, 11, 22].

In summary, teachers should provide learners with regular feedback that can encourage self-assessment and increase self-efficacy to develop self-regulation with clear standards of expected learning. In addition, teachers should provide learners with information that enables feedforward, such as prompting learners to resubmit assignments or increasing the difficulty of tasks in stages, so that learners can learn on their own initiative [22, 39]. We then consider the information to be included in the feedback to achieve such feedback.

What Information Should Be Included in Feedback Hattie and Timperley proposed a feedback model for learning effectiveness (in this paper, learning effectiveness refers to the usual effects on student achievement). This feedback model argues that teachers need to provide feedback to learners to answer the following questions [22].

- Where am I going?
- How am I going?
- Where to next?

In addition, Wiliam pointed out that teachers need to consider the following factors when evaluating learners [51].

- Where the learner needs to be
- Where the learner is right now
- How to get there

Both noted the need for items to clarify (1) what the learner's goal is, (2) what the learner is now doing, and (3) what the learner needs to do next. Hattie and Timperley noted that clarifying "what the learner needs to do next" is necessary for feedforward [22].

In addition, Brooks et al. suggested prompts on which feedback should be provided for each element [11]. In particular, they suggested that, depending on the learner's stage of learning, teachers should separate feedback about the task, process, and self-regulation.

From another perspective, the amount and complexity of feedback is one factor that teachers need to be careful about. Wisniewski et al., who conducted a meta-analysis of feedback studies, argued that the effects of feedback on learning are generally more effective when more information is contained in the feedback, although variability is significant [52]. However, it has also been pointed out that feedback that is too long may distract the learner and render the feedback useless [49] and that simple feedback is more effective for task-level feedback [22]. Furthermore, it has been shown that the amount of information in feedback may not affect learning effectiveness [5].

These inconsistent findings on the amount and complexity of feedback suggest that its effect is not solely dependent on information content and complexity. For example, Brooks et al. argued that the amount and complexity of information should be varied according to the learner's learning stage [11]. In this paper, we summarized the findings on the timing of feedback as other possible factors that may influence the effect of feedback.

When Feedback Should Be Provided Broadly, feedback can be categorized into two types: immediate and delayed. Immediate feedback refers to feedback that is given immediately after answering a question or test, whereas delayed feedback refers to feedback that is given hours or weeks apart compared to immediate feedback [49]. As we have mentioned the amount of information and complexity of feedback, the appropriate timing of feedback also has different effects on learning, depending on the study.

In a meta-analysis of studies focused on language learning, Kulik and Kulik showed that feedback timing can have different effects on different types of learning. They found that immediate feedback is effective in quizzes, which measure the comprehension of each learning topic, whereas delayed feedback was more effective in tests that assessed general knowledge and skills, and test scores were higher than immediate feedback [31]. In addition, Clariana et al. confirmed the effect of feedback in a high school class and found that delayed feedback was more effective when the difficulty level of the content was high, and immediate feedback was more effective when the difficulty level was low [16]. Similarly, Brooks et al. proposed varying the timing of the feedback depending on the learning stage. In particular, they demonstrated a strategy for delaying feedback as

learning evolved, and feedback was given to the self-regulation level [11]. Hattie and Timperley also argued that immediate feedback on a learner's learning activity is desirable when the achievement level is low [22].

In summary, the timing of appropriate feedback is related to the learning level. Immediate feedback is more effective in situations where the learning stage is not yet advanced and the learning content is simple, whereas delayed feedback is more effective as learning progresses and content becomes more complex. In distance PBL, which this study aims to support, individual learners are at different stages of learning as they progress, and it is necessary to appropriately adjust the timing for each learning situation.

Who Should Be Provided Feedback In PBL, it is important for teachers to know not only what kind of feedback is provided but also to whom to give feedback. For learners, too much feedback can be overwhelming, or they may not be able to receive the information in the feedback, and thus may not be able to benefit from the feedback [22, 36]. On the other hand, it is burdensome for teachers to constantly provide feedback to all learners. Therefore, a method for detecting learners who need feedback is required.

Sasaki et al. proposed a method to identify learners who need to be given feedback by analyzing video activity reports submitted by graduate students engaged in distance PBL and estimating learner engagement [45]. Activity reporting is a common part of learning activities [19] and is not particularly burdensome for learners. Engagement is also a measure of a person's mental state [21, 25, 47] and indicates how positively a person engages in a task when immersed in a single task [2]. Engagement has been defined in many fields, such as education and psychology [4]. In particular, the engagement proposed by Schaufeli et al. was originally used in the context of work but has also been shown to be significant for educational settings [46, 47]. Sasaki et al. analyzed video activity reports and found that there are relationships between the amount or length of filled or silent pauses in speech and engagement and that the inclusion of negative words in the content of activity reports is associated with lower learner engagement [45]. Decreased engagement indicates that the learners are not motivated to learn. As PBL requires learners to drive their learning activities [32], students who are less engaged are potentially less likely to succeed in PBL. Therefore, feedback and other support are required for students with declining engagement.

Sasaki et al. proposed a system that uses activity reports to detect learners whose engagement is declining at an early stage and notifies teachers [45]. Using this system, it is expected that teachers can identify potentially problematic learners based on their activity reports even if they do not always know how all learners are doing. In other words, teachers can easily determine to whom they should provide feedback.

3 Appropriate Feedback for Distance PBL

This study focuses on support distance PBL. In distance PBL, a teacher is usually in charge of multiple learners. In reality, however, teachers need to teach PBL and perform various other tasks, such as daily work and research activities. In other words, they cannot always pay attention to each learner and do not have an unlimited amount of time to provide feedback and other instructional activities. Therefore, a system that can easily provide feedback is expected to reduce the burden on teachers and, by extension, improve the quality of feedback, which is expected to have a positive effect on learners' learning. The purpose of this study is to organize insights into what appropriate feedback is in distance PBL.

A method for finding learners to whom feedback should be given has already been proposed by Sasaki et al. [45]. To develop a system for this research, it is necessary to clarify what kind of feedback should be provided to the learner. Therefore, in this paper, we first reviewed research on feedback in learning environments, investigated what kind of feedback is appropriate for learners, and summarized the results in the previous chapter. Then, based on the survey, we examine what appropriate feedback is in distance PBL.

One of the key principles of feedback is that teachers should clearly state the learning goals and standards for learners. Therefore, teachers are required to provide information about goals (Where am I going?) which are the prerequisites for PBL.

In PBL, the information provided in feedback changes according to the learner's learning stage. Therefore, both teachers and learners need to know how far their learning has progressed toward their goals. Information on the current learning stage is not only used to determine the content of feedback, but it is also expected to be used for self-assessment or self-adjustment level feedback, whereby the learner checks to see if there is any discrepancy with the current learning stage they expect to be at.

More specifically, an evaluation of what the learner is currently working on (How am I going?) is also needed. That is, teachers should show whether the learner's current learning activity is appropriate for the learning process, whether it is on the target, and whether it is the correct procedure. This feedback information works at both the process and self-regulation levels.

In addition, it is important to provide feedforward information which is what to do next (where to next?) to allow the learner to break the feedback loop and proceed to the next learning stage.

On the other hand, self-level feedback should not be added to the feedback, as it has been reported to have limited effects on learning and a negative impact on the learner.

The timing of the feedback depends on the learning stage. When learning has just begun, and the content of learning is simple, immediate feedback should be returned. When learning progresses and the learning content is detailed, delayed feedback should be returned.

Table 1. Prompt examples of feedback based on the learner stage

Learner Stage	*[Requirement 6]* Feedback Timing	*[Requirement 1]* **Feedback should show the learner's goal.**	*[Requirement 2]* **Feedback should show the learner's current stage.**	*[Requirement 3]* **Feedback should include the assessment for the learner's activity.**	*[Requirement 4]* **Feedback should include information about the next step.**
Basic	Immediate	• Your current goal of your study is ____.	• You are at the basic level of learning.	• Your work [has / has not] met the learning intention. • Your work [is / is not] what we are looking for.	• To achieve your work goal, you could ____. • [Adding / Removing] ____ would improve your work.
		• The key point in your task is ____.	• You are at the intermediate level of learning.	• Your understanding of your task is [sufficient / insufficient / ____]. • Your task progress [looks good / does not look good / ____].	• Thinking further about ____ could improve your work. • You could improve your ____ skills.
Advanced	Delayed	• Your goal is ____. • Are you focused on your purpose? • Could you monitor your work?	• You are at the advanced level of learning.	• Are you on track with your work? • Are you on track to achieving your goal?	• How could you improve your work? • Think what is the next step for your learning?

In summary, we show the following requirements for feedback in the system of feedback support for teachers in distance PBL, which this study will implement.

Feedback Requirement

1. Feedback should show the learner's goal.
2. Feedback should show the learner's current stage.
3. Feedback should include the assessment of the learner's activity.
4. Feedback should include information about the next step.
5. Feedback should NOT include evaluating the learners themselves.
6. The timing of providing feedback should vary according to the learner's stage.

Brooks et al. proposed three levels of feedback prompts according to the learner's learning stage [11]. Considering the prompts of Brooks et al., the aforementioned feedback requirements, and the context of distance PBL, this study

supports teachers providing feedback in distance PBL with the feedback timings and prompts shown in Table 1.

4 Design of a Feedback System to Support Distance Project-Based Learning

Based on the results of the survey, this chapter describes a feedback system to support teachers who provide feedback on distance PBL.

4.1 System Overview

We will develop the system shown in Figure 1, as proposed by Sasaki et al. This system uses video activity reports submitted by learners. The system analyzes the activity report, and when it detects that the learner's engagement is decreasing, the system notifies the teacher that the learner's engagement is decreasing and encourages them to give feedback. Feedback is provided by text. The system automatically suggests the text of the feedback and prepares a form of feedback, aiming to reduce the burden on teachers to prepare the feedback text.

4.2 Functions

The system to support teacher feedback that this study will build will have the following functions.

Analyzing the Submitted Report Learners regularly submit video activity reports while engaging in PBL. The video activity reports are analyzed based on Sasaki et al.'s report to estimate whether learner engagement is declining [45]. If learner engagement is estimated to decline, a notification is sent to teachers, encouraging them to give feedback to the learner in question. The notification will include a link to a form designed to simplify the feedback process.

Preparing a Feedback Form Figure 2 illustrates the method of creating feedback to be presented to teachers. On the screen (a), the video activity report and transcribed report submitted by the learner, whose engagement is estimated to be declining and who is presumed to need feedback, are displayed. If necessary, past activity reports can also be viewed. While viewing the video activity report and the content of the report, the teacher inputs the content into the feedback form in accordance with the feedback requirements. The form should allow the teacher to select the content to be entered from options based on the learning stage of the last time the feedback was given to the learner, refer to information from the previous feedback, or write the content freely. When inputting information, the system moves to the screen (b).

On the screen (b), feedback is reviewed before it is sent to the learner. Modify it if necessary and determine the feedback text.

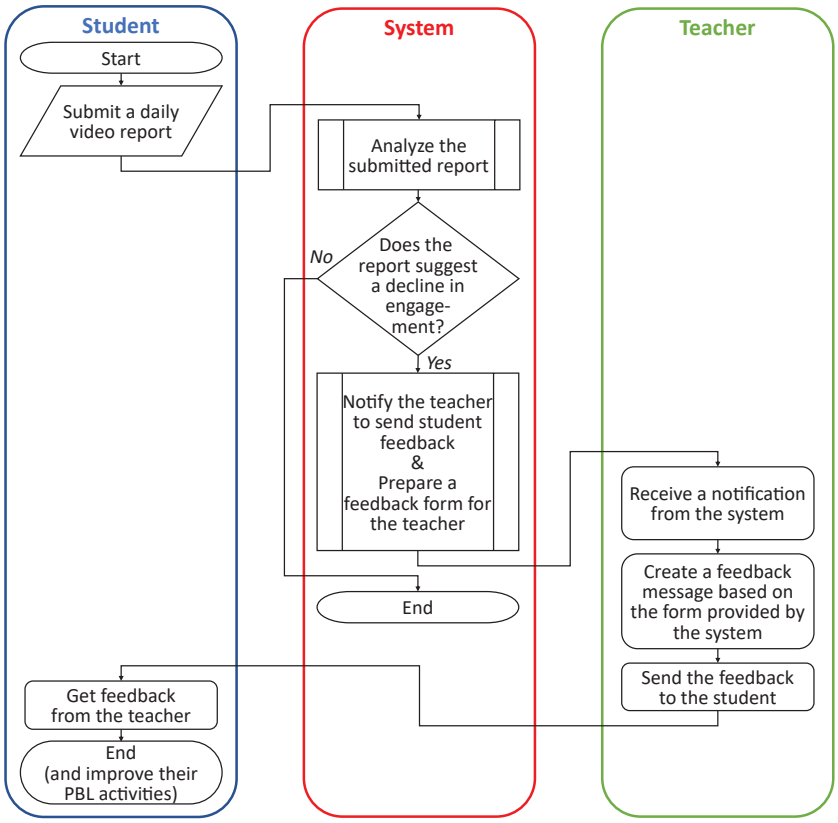


Fig. 1. Feedback system based on [45]

Delaying Sending Feedback The earlier the learning stage, the more immediate feedback is preferred. However, if learning is advanced, feedback should be delayed. Therefore, the system delays the timing of sending the generated feedback according to the learning stage entered on the screen (a) of Figure 2.

4.3 Future Work

Currently, the system has not been implemented, and further study is needed on the implementation of the system, especially on the specific delay time in the delayed feedback. We will then evaluate whether the use of this system will reduce the burden of feedback on teachers, and whether learners who receive feedback will be able to proceed smoothly with PBL.

5 Conclusion

This study focused on distance PBL. In PBL, appropriate feedback from teachers is necessary for learners to proceed smoothly with learning activities. However,

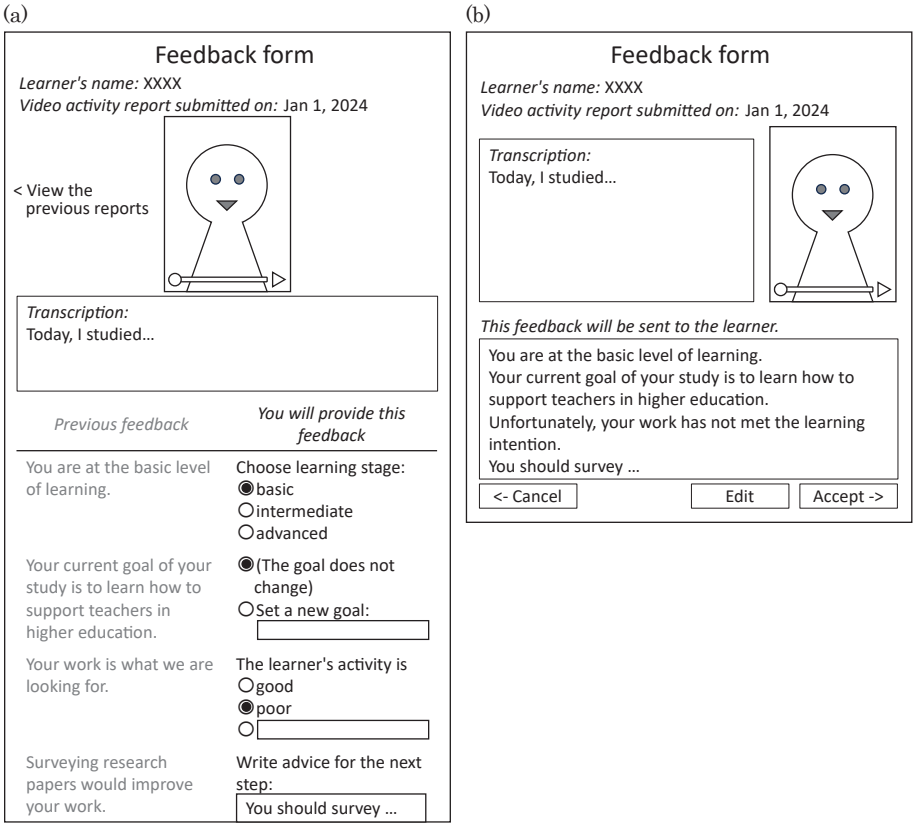


Fig. 2. Illustration of a feedback form. Confirm the video activity report and input feedback information on screen (a). Confirm the feedback content on screen (b).

in distance PBL, it is difficult for teachers to track multiple learners at all times. Therefore, this study aimed to construct a system that can easily provide appropriate feedback to support teachers in distance PBL. In previous studies, a method to detect which kind of learner should be given feedback using activity reports submitted by learners has been considered [45]. However, how feedback should be provided in detail has not been sufficiently studied. In this paper, we surveyed previous studies on feedback to organize findings on appropriate feedback in distance PBL. From the survey results, we summarized the six feedback requirements in distance PBL and indicated when and what prompt feedback should be provided to learners. We also presented an overview of a system to assist teachers in providing feedback that meets the feedback requirements.

References

1. Adnan, M., Anwar, K.: Online learning amid the covid-19 pandemic: Students' perspectives. Online Submission **2**(1), 45–51 (2020)
2. Afflerbach, P., Harrison, C.: What is engagement, how is it different from motivation, and how can i promote it? Journal of Adolescent & Adult Literacy **61**(2), 217–220 (2017)
3. Almulla, M. A.: The effectiveness of the project-based learning (pbl) approach as a way to engage students in learning. SAGE Open **10**(3), 2158244020938702 (2020)
4. Azevedo, R.: Defining and measuring engagement and learning in science: Conceptual, theoretical, methodological, and analytical issues. Educational Psychologist **50**(1), 84–94 (2015)
5. Azevedo, R., Bernard, R. M.: A meta-analysis of the effects of feedback in computer-based instruction. Journal of Educational Computing Research **13**(2), 111–127 (1995)
6. Bandura, A.: Self-efficacy: Toward a unifying theory of behavioral change. Psychological Review **84**(2), 191–215 (1977)
7. Bangert-Drowns, R. L., Kulik, C.-L. C., Kulik, J. A., Morgan, M.: The instructional effect of feedback in test-like events. Review of Educational Research **61**(2), 213–238 (1991)
8. Black, P., Wiliam, D.: Assessment and classroom learning. Assessment in Education: Principles, Policy & Practice **5**(1), 7–74 (1998)
9. Black, P., Wiliam, D.: Developing the theory of formative assessment. Educational Assessment, Evaluation and Accountability (formerly: Journal of Personnel Evaluation in Education) **21**, 5–31 (2009)
10. Boud, D., Molloy, E.: Rethinking models of feedback for learning: The challenge of design. Assessment & Evaluation in Higher Education **38**(6), 698–712 (2013)
11. Brooks, C., Carroll, A., Gillies, R. M., Hattie, J.: A matrix of feedback for learning. Australian Journal of Teacher Education (Online) **44**(4), 14–32 (2019)
12. Brophy, J.: Teacher praise: A functional analysis. Review of Educational Research **51**(1), 5–32 (1981)
13. Brown, G. T., Peterson, E. R., Yao, E. S.: Student conceptions of feedback: Impact on self-regulation, self-efficacy, and academic achievement. British Journal of Educational Psychology **86**(4), 606–629 (2016)
14. Cavalcanti, A. P., Diego, A., Mello, R. F., Mangaroska, K., Nascimento, A., Freitas, F., Gašević, D.: How good is my feedback? a content analysis of written feedback. In: Proceedings of the Tenth International Conference on Learning Analytics & Knowledge. p. 428–437. LAK '20, Association for Computing Machinery, New York, NY, USA (2020)
15. Chen, Z., Cao, H., Deng, Y., Gao, X., Piao, J., Xu, F., Zhang, Y., Li, Y.: 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. pp. 1–16 (2021)
16. Clariana, R. B., Wagner, D., Roher Murphy, L. C.: Applying a connectionist description of feedback timing. Educational Technology Research and Development **48**(3), 5–22 (2000)
17. Cojocariu, V.-M., Lazar, I., Nedeff, V., Lazar, G.: Swot anlysis of e-learning educational services from the perspective of their beneficiaries. Procedia-Social and Behavioral Sciences **116**, 1999–2003 (2014)

18. Earley, P. C., Northcraft, G. B., Lee, C., Lituchy, T. R.: Impact of process and outcome feedback on the relation of goal setting to task performance. Academy of Management Journal **33**(1), 87–105 (1990)
19. Etkina, E., Harper, K. A.: Weekly reports: Student reflections on learning. Journal of College Science Teaching **31**(7), 476–480 (2002)
20. Glassey, R.: Developing feedback analytics: Discovering feedback patterns in an introductory course. In: Proceedings of the ACM Conference on Global Computing Education. p. 37–43. CompEd '19, Association for Computing Machinery, New York, NY, USA (2019)
21. Harter, J. K., Schmidt, F. L., Hayes, T. L.: Business-unit-level relationship between employee satisfaction, employee engagement, and business outcomes: A meta-analysis. Journal of Applied Psychology **87**(2), 268–279 (2002)
22. Hattie, J., Timperley, H.: The power of feedback. Review of Educational Research **77**(1), 81–112 (2007)
23. Hounsell, D., McCune, V., Hounsell, J., Litjens, J.: The quality of guidance and feedback to students. Higher Education Research & Development **27**(1), 55–67 (2008)
24. Johnson, B., Zimmermann, T., Bird, C.: The effect of work environments on productivity and satisfaction of software engineers. IEEE Transactions on Software Engineering **47**, 736–757 (2021)
25. Kahn, W. A.: Psychological conditions of personal engagement and disengagement at work. Academy of Management Journal **33**(4), 692–724 (1990)
26. Karabenick, S. A., Knapp, J. R.: Relationship of academic help seeking to the use of learning strategies and other instrumental achievement behavior in college students. Journal of Educational Psychology **83**(2), 221–230 (1991)
27. Kluger, A. N., DeNisi, A.: The effects of feedback interventions on performance: A historical review, a meta-analysis, and a preliminary feedback intervention theory. Psychological Bulletin **119**(2), 254–284 (1996)
28. Kluger, A. N., DeNisi, A.: Feedback interventions: Toward the understanding of a double-edged sword. Current Directions in Psychological Science **7**(3), 67–72 (1998)
29. Kulhavy, R. W.: Feedback in written instruction. Review of Educational Research **47**(2), 211–232 (1977)
30. Kulhavy, R. W., Stock, W. A.: Feedback in written instruction: The place of response certitude. Educational Psychology Review **1**, 279–308 (1989)
31. Kulik, J. A., Kulik, C.-L. C.: Timing of feedback and verbal learning. Review of Educational Research **58**(1), 79–97 (1988)
32. Markham, T.: Project based learning a bridge just far enough. Teacher Librarian **39**(2), 38–42 (2011)
33. Mory, E. H.: The use of informational feedback in instruction: Implications for future research. Educational Technology Research and Development **40**, 5–20 (1992)
34. Mory, E. H.: Feedback research revisited. In: Handbook of Research on Educational Communications and Technology, pp. 738–776. Routledge (2013)
35. Mulki, J. P., Bardhi, F., Lassk, F. G., Nanavaty-Dahl, J.: Set up remote workers to thrive. MIT Sloan Management Review **51**(1), 63 (2009)
36. Naeger, D. M., Jen, A., Ahearn, B., Webb, E. M.: Effectively acquiring and using feedback. Journal of the American College of Radiology: JACR **12**(12 Pt A), 1320–1323 (2015)
37. Narciss, S.: Feedback strategies for interactive learning tasks. In: Handbook of Research on Educational Communications and Technology, pp. 125–143. Routledge (2008)

38. Newman, R. S., Schwager, M. T.: Students' perceptions of the teacher and classmates in relation to reported help seeking in math class. The Elementary School Journal **94**(1), 3–17 (1993)
39. Nicol, D. J., Macfarlane-Dick, D.: Formative assessment and self-regulated learning: A model and seven principles of good feedback practice. Studies in Higher Education **31**(2), 199–218 (2006)
40. Paris, S. G., Winograd, P.: Promoting metacognition and motivation of exceptional children. Remedial and Special Education **11**(6), 7–15 (1990)
41. Riordan, T., Loacker, G.: Collaborative and systemic assessment of student learning: From principles to practice. In: Assessment, Learning and Judgement in Higher Education, pp. 1–18. Springer (2008)
42. Ryan, A. M., Gheen, M. H., Midgley, C.: Why do some students avoid asking for help? an examination of the interplay among students' academic efficacy, teachers' social–emotional role, and the classroom goal structure. Journal of Educational Psychology **90**(3), 528–535 (1998)
43. Ryan, A. M., Pintrich, P. R., Midgley, C.: Avoiding seeking help in the classroom: Who and why? Educational Psychology Review **13**, 93–114 (2001)
44. Sadler, D. R.: Formative assessment and the design of instructional systems. Instructional Science **18**(2), 119–144 (1989)
45. Sasaki, K., He, Z., Inoue, T.: Using video activity reports to support remote project-based learning. JUCS - Journal of Universal Computer Science **29**(11), 1336–1360 (2023)
46. Schaufeli, W. B., Martinez, I. M., Pinto, A. M., Salanova, M., Bakker, A. B.: Burnout and engagement in university students: A cross-national study. Journal of Cross-Cultural Psychology **33**(5), 464–481 (2002)
47. Schaufeli, W. B., Salanova, M., González-Romá, V., Bakker, A. B.: The measurement of engagement and burnout: A two sample confirmatory factor analytic approach. Journal of Happiness Studies **3**(1), 71–92 (2002)
48. Schunk, D. H.: Self-efficacy and achievement behaviors. Educational Psychology Review **1**, 173–208 (1989)
49. Shute, V. J.: Focus on formative feedback. Review of Educational Research **78**(1), 153–189 (2008)
50. Thompson, T., Richardson, A.: Self-handicapping status, claimed self-handicaps and reduced practice effort following success and failure feedback. British Journal of Educational Psychology **71**(1), 151–170 (2001)
51. Wiliam, D.: Assessment: The bridge between teaching and learning. Voices from the Middle **21**(2), 15–20 (2013)
52. Wisniewski, B., Zierer, K., Hattie, J.: The power of feedback revisited: A meta-analysis of educational feedback research. Frontiers in Psychology **10**, 487662 (2020)

Orientation-Dependent Cord Length Distribution Functions of Bounded Convex Domains ⋆

N. G. Aharonyan and V. K. Ohanyan[0000−0001−7029−2385]

American University of Armenia and Yerevan State University
victo@aua.am, victoohanyan@ysu.am, narine78@ysu.am

Abstract. In the last century German mathematician W. Blaschke formulated the problem of investigation of bounded convex domains in the plane using probabilistic methods. In particular, the problem of recognition bounded convex domains $\mathbb{D}$ by chord length distribution function (or density function).

Keywords: Stochastic and Integral geometry; Chord length distribution.

1 Introduction

The present paper continues the investigations begun in [11]. Random lines generate chords of random length in convex domain $\mathbb{D}$. The corresponding distribution (or density) function is called the chord length distribution function that we denote by $F(y)$ (or chord length density function $f(y)$). The form of the length density function is related to certain features of the corresponding figures. Poles of this function are related to parallel pieces of the contour and the form of $f(y)$ for y close to its maximum is essentially related to smaller details of the contour. The determination of the chord length distribution function has a long tradition of application to collections of bounded convex bodies forming structures in metal and crystallography. The series of formulae for chord length distribution functions may be of use in finding suitable models when empirical distribution functions are given (see [9]).

In the initial stage of investigation mathematicians tried to find explicit expressions of the chord length distribution (or density) functions for concrete domains $\mathbb{D}$ in the terms of elementary functions. Till recently explicit expressions for the chord length distribution functions have been known in the case when $\mathbb{D}$ is a disc, a regular triangle (see [5]) and a rectangle (see [6]). These results have been obtained using the definition of chord length distribution function for a domain $\mathbb{D}$.

⋆ The investigation of the second author is done with partial support by the Mathematical Studies Center at Yerevan State University.

A family of identities primarily associated with isoperimetric inequalities for planar convex domains was discovered by A. Pleijel (see [4]). Using combinatorial principles in integral geometry R. V. Ambartzumian proved these identities (see [1], [2]) and pointed out (see [2], section 6.10) that classical Pleijel identities can be used for the finding chord length distribution function for the polygons which have no parallel sides.

Using delta-formalism in Pleijel identities we have obtained explicit expression for the chord length distribution function for any regular polygon (see [10]). In the particular cases of a regular triangle, a square, a regular pentagon and a regular hexagon our formula for the chord length distribution function coincides with formulas available in the literature (see [5], [6], [7] and [8]) for $n = 3, 4, 5, 6$ correspondingly.

2 Chord Length Distribution Functions

Let $\mathbb{D}$ be a convex bounded polygon in the plane and a_1,..., a_n be sides of $\mathbb{D}$. The so–called Pleijel identity for $\mathbb{D}$ is as follows:

$$\int_{[\mathbb{D}]} f(|\chi(g)|)\, dg = \int_{\mathbb{G}} f'(|\chi|)\cdot|\chi|\, \cot\alpha_1\, \cot\alpha_2\, dg + \sum_{i=1}^{n}\int_0^{|a_i|} f(u)\, du, \qquad (1)$$

where $f(x)$ is a function with continuous first derivative $f'(x)$, α_1 and α_2 are the angles between $\partial\mathbb{D}$ and g at the endpoints of $\chi(g) = g \cap \mathbb{D}$ which lie in one half-plane with respect to the inside of $\mathbb{D}$, $|a_i|$ is the length of a_i, $i = 1, ..., n$.

Let $\mathbb{G}$ be the space of all lines g in the Euclidean plane $\mathbb{R}^2$, (p, φ) = the polar coordinates of the foot of the perpendicular to g from the origin O, be standard coordinates for a line $g \in \mathbb{G}$. Then

$$[\mathbb{D}] = \{g \in \mathbb{G} :\ g \cap \mathbb{D} \neq \emptyset\} = \bigcup_{i<j}([a_i] \cap [a_j])$$

where $[a_i] \cap [a_j]$ is the set of lines hitting both sides a_i and a_j of $\mathbb{D}$.

Let b_{ij} be the distance between the parallel segments a_i and a_j (i.e. the distance between the lines containing a_i and a_j). Further, $h(\varphi) \neq b_{ij}$ is the height of the maximal parallelogram with two sides equal to $\chi(\varphi) = g(\varphi) \cap \mathbb{D}$, $g(\varphi) \in [a_i] \cap [a_j]$ ($g(\varphi)$ is a line with φ-direction), and the other two sides lie on the parallel sides a_i and a_j,

$$h(\varphi_\chi) = h\left(\arccos\frac{b_{ij}}{|\chi(\varphi)|}\right) + h\left(2\,\pi - \arccos\frac{b_{ij}}{|\chi(\varphi)|}\right).$$

Hence , $h(\cdot) = 0$ if the parallelogram is empty.
For the value of φ such that $|\chi(\varphi)| = y$ we have $h(\varphi_y) = h(\varphi_\chi)$.

We obtain

$$F(y) = 1 - \frac{1}{\sum_{i=1}^{n} |a_i|} \left[\sum_{i<j}^{I} \frac{y}{\sin \gamma_{ij}} \int_{\Phi_{ij}(y)} \sin\varphi \, \sin(\gamma_{ij} - \varphi) \, d\varphi - \right.$$
$$\left. - \sum_{i<j}^{II} I_{ij}(y) \, h(\varphi_y) \, \frac{\sqrt{y^2 - b_{ij}^2}}{b_{ij}} + \sum_{i=1}^{n} (|a_i| - y)^+ \right], \qquad (2)$$

where $\sum^I$ is over all nonparallel pairs of segments a_i, $a_j \subset \partial\mathbb{D}$ and $\sum^{II}$ is over all parallel pairs of segments a_i and $a_j \subset \partial\mathbb{D}$.

In the case where $\partial\mathbb{D}$ contains no pairs of parallel sides, formula (2) coincides with the expression given by R. V. Ambartzumian in [2], page 158.

It follows from (2) that to find distribution function $F(y)$ we have to calculate integrals of the form

$$\frac{1}{\sin\gamma} \int_{\Phi(y)} \sin\varphi \, \sin(\gamma - \varphi) \, d\varphi$$

for any two nonparallel segments a and b ($a \le b$) with the angle γ between a and b (or their continuations) and also calculate the second sum in (2) (for pairs of parallel sides). Here $\Phi(y)$ is

$$\Phi(y) = \{\varphi : \text{a chord joining } a \text{ and } b \text{ exists with direction } \varphi \text{ and length } y\}.$$

For the proof of (2) see [10] and [7].

In the last years have been introduced the notion of orientation-dependent chord length distribution function $F_\phi(y)$, while F(y) is called mixed orientation distribution function.

In the paper [8] an explicit formula for orientation-dependent chord length distribution function for any bounded convex domains $\mathbb{D}$ have been obtained. Using delta-formalism we have obtained orientation-dependent chord length distribution function (see [8]).

3 Chord Length Distribution Functions for Regular Polygons

In the classical Pleijel identities integration is over the measure in the space **G** of lines which is invariant with respect to the all Euclidean motions (see [1], [2]). In the paper [8] generalized Pleijel identities for any locally-finite, bundleless measure in the space **G** have been proved. The following identity

$$\int_{\mathbb{G}} |\chi(g)|^n \, m(dg) = \frac{n}{2} \int_{\mathbb{G}} |\chi(g)|^{n-1} \, m([\chi(g)]) \, dg -$$
$$- \frac{n(n-1)}{4} \int_{(\partial D)^2} \chi_{12}^{n-2} m([\chi_{12}]) \, \cos\alpha_1 \, \cos\alpha_2 \, dl_1 \, dl_2 \qquad (3)$$

is called generalized Pleijel identity.

In particular, if $m(dg) = dg$, i.e. $m([\chi(g)]) = 2 \cdot |\chi(g)|$ we obtain classical Pleijel identity.

This identity is applied to find the so-called orientation-dependent chord length distribution functions for bounded convex domains:

$$b(\phi_0) \cdot [1 - F_{\phi_0}(y)] = \frac{1}{2} \int_{[\mathbb{D}]} \delta(|\chi(g)| - y)\, |\chi(g)|\, |\sin(\phi - \phi_0)|\, dg -$$

$$-\frac{1}{2} \int_{[\mathbb{D}]} \delta'(|\chi(g)| - y) \cdot |\chi(g)|^2\, |\sin(\phi - \phi_0)|\, \cot\alpha_1\, \cot\alpha_2\, dg, \qquad (4)$$

where $b(\phi)$ is the breadth function (ϕ is a direction), $\delta(y)$ is the Dirac's δ–function concentrated at y, α_1 and α_2 are the angles between $\partial\mathbb{D}$ and the line g at the endpoints of $\chi(g) = g \cap \mathbb{D}$ which lie in one half-plane with respect to the inside of $\mathbb{D}$, while $F_\phi(y)$ is the orientation-dependent distribution function at ϕ

$$F_{\phi_0}(y) = \frac{1}{b(\phi_0)}\, \mathcal{L}_1\{p :\ |\chi(p, \phi_0)| < y\},$$

where $\mathcal{L}_1$ is one dimensional Lebesgue measure.

Proposition ([12]). Let **D** be a convex planar polygon which has m pairs of parallel sides (a_{i_1}, a_{j_1}),...,(a_{i_m}, a_{j_m}). The distances of the parallel lines which carry these segments are d_1,...,d_m, respectively, and $\pi a_{i_k} \cap \pi a_{j_k}$ denotes the length of the intersection of the orthogonal projections of both segments onto one of the carrying lines, $k = 1, ..., m$. Then for $k \in \{1, ..., m\}$ for which $\pi a_{i_k} \cap \pi a_{j_k} > 0$, the chord length density function has a discontinuity at d_k, and the limit from above at d_k is infinite.

Assume now to have the information about the distribution of the chord length not in the "completely mixed" form, but separated direction by direction. One can show that the problem of determining a body from this data, gave a positive answer using "orientation dependent" chord length distribution, when $\mathbb{D}$ is a planar convex polygon (see [9]). A set of four directions whose slopes have a transcendental cross ratio will ensure that the corresponding parallel X-rays determine each planar convex domain (see [9]). The following problem arise.

Does there exist a finite set of directions $V = \{\phi_1, ..., \phi_m\}$ such that the corresponding set of "orientation-dependent" chord length distribution functions $F_{\phi_1}(y)$, ..., $F_{\phi_m}(y)$ determine a bounded convex domain uniquely.

This question received negative answer, because it is possible to construct two non-congruent triangles that have the same chord length distribution function for a fixed set of m directions, where m is a natural (see [11]). The question arises whether it is possible to find a subclass of convex bodies, where it is possible to reconstruct a body from the values of $F_D(u, x)$ for a finite set of directions.

References

1. Ambartzumian, R. V.: Combinatorial Integral Geometry with Applications to Mathematical Stereology. John Wiley and Sons, Chichester (1982)

2. Ambartzumian, R. V.: Factorization Calculus and Geometric Probability. Cambridge University Press, Cambridge (1990)
3. Ohanyan, V. K.: Combinatorial principles in Stochastic Geometry: A Review. Journal of Contemporary Mathematical Analysis **43**(1), 44–60 (2008)
4. Pleijel, A.: Zwei kurze Beweise der isoperimetrischen Ungleichung. Archiv Math. **7**, 317–319 (1956)
5. Sulanke, R.: Die Verteilung der Sehnenlängen an ebenen und räumlichen Figuren. Math. Nachr. **23**, 51–74 (1961)
6. Gille, W.: The chord length distribution of parallelepipeds with their limiting cases. Exp. Techn. Phys. **36**, 197–208 (1988)
7. Aharonyan, N. G. and Ohanyan, V. K.: Chord length distribution functions for polygons. Journal of Contemporary Mathematical Analysis **40**(4), 43–56 (2005)
8. Aharonyan, N. G.: Generalized Pleijel Identity. Journal of Contemporary Mathematical Analysis **43**(5), 3–12 (2008)
9. Gardner, R. J.: Geometric Tomography. 2nd edn. Cambridge University Press, Cambridge, UK, New York (2006)
10. Ohanyan, V. K. and Harutyunyan, H.: Chord length distribution functions for regular polygons. Advances in Applied Probability **41**, 358–366 (2009)
11. Ohanyan, V. K.: Recognition of convex bodies by tomographic methods, In Proceedings of 3rd CODASSCA workshop, Data Science, Human-Centered Computing, and Intelligent Technologies, Logos Verlag, Berlin (2022)
12. Ohanyan, V. K. Martirosyan, D. M.: On intersection probabilities of four lines inside a planar convex domain. Journal of Advances in Applied Probability **60**(2), 416–430 (2023)
13. Martirosyan, D. M., and Ohanyan, V. K.: On the Euclidean Distance Between Two Gaussian Points and the Normal Covariogram of R^d. Journal of Contemporary Mathematical Analysis **59**(1), 38–46 (2024)

Assessing Glaucoma Online Tools

Nelson Baloian[1][0000-0003-1608-6454] and Wolfram Luther[0000-0002-1245-7628]

[1] Department of Computer Science, University of Chile, Santiago, Chile
[2] Department of Computer Science, University of Duisburg-Essen, Germany

wolfram.luther@uni-due.de, nbaloian@dcc.uchile.cl

Abstract. In a recent publication, we presented online tools for computing the familial risk of stroke, for the occurrence of pathogenic variants in the BRCA1 or BRCA2 genes with impact on early breast and ovarian cancer disease, and for low- and high-stage prostate cancer [2, 3]. The forms collect information about the individuals, their specific disease patterns, medical examination results, and the lifestyle of the proband and his/her relatives. Furthermore, changes to the examination methods and redefinition of risk classes, data and model quality, as well as cross-cutting issues such as uncertainty and usability have been addressed. In this paper, we present various approaches, continuous or simple score-based risk models for estimating the 5-year risk that an individual with ocular hypertension will develop Primary Open Angle Glaucoma (POAG), the leading global cause of irreversible blindness. Finally, a customizable Dempster-Shafer (DS) risk assessment model is derived.

Keywords: Primary Open Angle Glaucoma Risk Calculator, Quality Metrics, Dempster-Shafer Risk Model Assessment.

1 Introduction

At least 2.2 billion people have a near or distance vision impairment. In almost half of these cases, the visual impairment could have been prevented or be remedied by preventive measures: early detection, periodical monitoring and appropriate treatment are crucial to avoid progressive and irreversible vision loss (within the next 5 years). But depending on the medical care available in their countries, often far too few people with distance vision problems due to refractive errors or a cataract have access to appropriate treatment [16].

The main cause of vision impairment is presbyopia, and to a far lesser extent:

- refractive errors
- cataract
- diabetic retinopathy
- glaucoma (Prevalence up to 111.8 million people worldwide by 2040, [12])
- age-related macular degeneration.

Glaucoma represents a degenerative optic neuropathy characterized by the progressive degeneration of retinal ganglion cells and the retinal nerve fiber layer, which leads to corresponding visual field defects [17]. There are four principal types: Primary Open-Angle Glaucoma (POAG) accounts for at least 90% of all cases–the angle between the iris and cornea remains open and drainage does not work properly–, followed by Closed-Angle, Congenital, a rare form of glaucoma in infancy, and Secondary Glaucoma. In addition to increased intraocular pressure, advanced age, patient's origin and

the presence of the disease in close relatives, other important factors exist for the onset of vision loss and damage to the optic nerve [13].

There are four and more stages of glaucoma symptoms and various classification systems (https://www.healthline.com/health/eye-health/stages-of-glaucoma). Special pathological variants in genes are responsible for some forms of premature glaucoma. Measurement and assessment of visual acuity, visual field screening, contrast sensitivity, glare testing and color vision are important criteria for ocular health assessment and indicators for existing or prospective impairments.

Online glaucoma calculators, which are presented in this article, are based on a mathematical model for estimating the 5- or *n*-year risk that an individual with ocular hypertension develops a POAG [5], [7], [9]. Several characteristics of an individual are combined in order to compose a risk metric. These include patient's age (y), means (of several measures) of intraocular pressure (IOP, mmHg), central corneal thickness (CCT, mm), vertical cup-to-disc ratio (VCD), visual field pattern standard deviation (PSD, dB) and other parameters as (corrected Octopus) loss variance (OLV, dB)–the local heterogeneity of a visual field defect are used in a composite metric.

Risk increases for persons with older age, higher IOP, larger VCD, thinner CCT, higher PSD, LV and an additional existing diabetes disease, which appears to have a protective effect against POAG. Common risk metrics use 2–6 of the most relevant parameters, but they should consider that age-adjusted prevalence rates of POAG strongly depend on the origin and patient's family history. Two parameters describing visual quality can also be used for validation purposes.

A comparison of the results of risk calculators and related studies is only possible if similar patient groups are medically treated according to the same procedures and similar protocols and if the measurements use the same procedures and model parameters and concern overlapping time periods.

For example, measurements are undertaken for both eyes several times in certain time periods under defined conditions using the same procedures, and a statistical model adjustment and calibration have been carried out.

Epidemic uncertainty in the results occurs if measured values are missing due to the absence of test subjects or if personal data is missing or incorrectly collected.

2 Prevention Criteria and Online Risk Tools

In this section, we present three point-based risk calculators that use 3–5 input variables to assign patients to one of three risk groups. Low, medium and high risk are characterized using percentage intervals, and the percentage probability of developing POAG within the next five years is indicated. At the same time, recommendations are made depending on the risk class, ranging from counseling and check-ups to targeted treatments and therapies that are carried out according to a specific schedule.

2.1 Laroche Glaucoma Risk Calculator (LGRC)

Based on three measurements per eye and averaging the proband's integer values of age with $100>y>40$, IOT >12 and $375\leq CCT<625$, the Laroche low-cost calculator [7] determines a result of the point score, identifies Glaucoma patients and recommends measures for the prevention and treatment of the reduced vision corresponding to the risk class and the recommendations of national and international care organizations. A

score of less than 6 points means low risk, and reevaluation is recommended in one year. A result of 6 or more points signifies high risk and recommends immediate, complete ophthalmic evaluation and adequate treatment.

The study describes in detail the demographic data of the glaucoma and control cohorts, including statistical means and standard deviations.

LGRC: Total Score (TS):= ⌈(age-40)/10⌉ + min(⌈(IOP-12)/3⌉, ⌈4+(IOP-25)/4⌉) − ⌈(CCT-550)/25⌉, Low risk: TS≤5 points, high risk: TS ≥6 points.

An asymmetry in the interval sizes in the original calculator was corrected. Limitations are found under https://irp.cdn-website.com/4d783a5d/files/uploaded/Laroche-Glaucoma-Calculator-PDF-KRG-DL-Updated.pdf. There is no guarantee that the model predicts the progression of an established disease or the development of a visual disability from sight defects, (diabetic)-retinopathy, cataracts and other diseases.

The score is validated based on optical coherence tomography (OCT), cup-to-disc ratio, and/or visual field anomalies associated with structural glaucoma damage. The mean values in the total sample are age=65.4, IOP=21.2, and CCT=531.9. For validation, the study uses results from the Ocular Hypertension Treatment Study (OHTS) and the European Glaucoma Prevention Study (EGPS).

The proposed glaucoma calculator and its validation have some disadvantages:

- Only a small subject group of 104 patients (54 with glaucoma and 50 controls) was considered, without a family history or calibration of the calculator using the specific mean values of the groups. Finally, there are considerable rounding effects.
- High uncertainty: Glaucoma risk score of a patient in the control group with mean values of parameters is 5 points
- Discrete point scores lead to fluctuations of up to three points in the output value for small changes in the input variables
- Ex.: Patient with 65y, CCT=532 mm, IOP 21.2 mmHg, VCD=0.7, PSD=2.0 dB. LGRC score: 6 pts., high risk; OHTS (section 2.2): 12 pts., 20%; [1] continuous POAG risk calculator: 26.9%; Our DS-like model (DSM) introduced in section 2.5 provides: 21%.

2.2 Ocular Hypertension Treatment Study (OHTS) Calculator

The Ocular Hypertension Treatment Study and the European Glaucoma Prevention Study, which studied the treatment of ocular hypertension, provided the basis for a score-based and a continuous version of the Glaucoma Risk Calculator.

Points	0	+1	+2	+3	+4
Age, y	30-44	45-54	55-64	65-74	≥75
Mean IOP	<22	22 to <24	24 to <26	26 to <28	≥28
Mean CCT	>600	576-600	551-575	526-550	≤525
Mean VCD	<0.3	0.3 to <0.4	0.4 to <0.5	0.5 to <0.6	≥0.6
Visual field					
Mean PSD	<1.8	1.8 to <2.0	2.0 to <2.4	2.4 to <2.8	≥2.8
Mean OLV	<3.24	3.24 to <4.0	4.0 to <5.76	5.76 to <7.84	≥7.84

Both were originally published by Michael A. Kass, MD, professor of ophthalmology and chairman of the Department of Ophthalmology and Visual Sciences at Washington University in St. Louis, and his team, cf. https://www.mdcalc.com/calc/ 10025/ocular-hypertension-treatment-study-ohts-calculator and [1].

The calculators require the following patient information: Age (y), several IOP measurements on both eyes and averaging, CCT using an ultrasound pachymeter, pattern standard deviation using any of the following Humphrey full threshold 30-2 or 24-2, SITA standard 30-2 or 24-2, or loss variance from Octopus 32-2 [14].

After evaluating the scores based on the patient's examination results, they are assigned to the following risk classes: Low <7 pts: 10% risk, intermediate<13 pts: 20% risk, and high ≥13 pts: ≥33% risk. Recommendation: counseling on risks and initiation of treatment; benefits of treatment versus close observation; and percentage risk of developing POAG within 5 years.

To validate the POAG risk prevalence, the authors of [5] followed a patient group of 1636 participants who underwent standardized ophthalmologic treatment (mean age 55.4 years, 56.9% women, 1138 white people, 407 black/African American, and 515 subjects dropped out of the study due to death) for up to 20 years and compared individual data on the progression of visual loss with the subjects POAG predictions of the tool. The 20-year cumulative incidence of POAG was 46% in one or both eyes, and the one of visual field loss was 25% after adjusting for exposure time [10].

2.3 WUSM POAG Risk Calculator (OHTS/EGPS)

This risk calculator has a point-system model and a continuous method and estimates the 5-year risk of developing POAG. The following patient'sdata are entered in a screen mask: age (30-80 years), IOP (20-32 mmHg), VCD (0-0.8) by contour, CCT (450-650 mm), PSD the corrected loss variance (0.5-3 Humphrey, Octopus perimeters), all ocular data as standardized measurements (cf. Figure, 1, https://ohts.wustl.edu/risk/ and link Use Continuous Method (pdf))

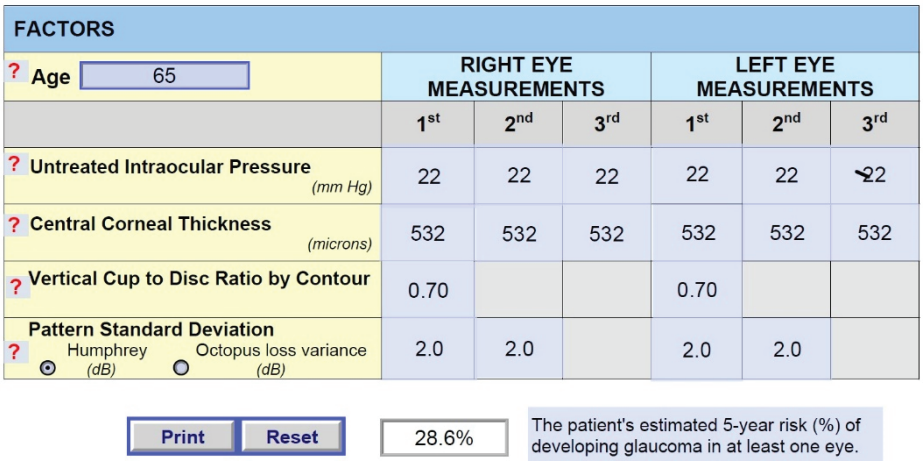


Fig. 1. The continuous method WPOAG for estimating the 5-year risk R of developing POAG $R= 1-0.911^{\exp(0.2792\cdot((age-56)/12 + (IOP-25)/3) +0.75566\cdot(573-CCT)/42 +0.6262\cdot((PSD-1.9) + 10(VCD-3.9)/3.5))}$

Pattern standard deviation/corrected loss variance needs two measurements per eye, using either Humphrey or Octopus perimeters. Octopus Corrected Loss Variance measurements will be converted internally to Humphrey PSD equivalent. In [4], the authors confirm these prediction parameters in the two cohorts studied—the OHTS observation group and the EGPS placebo group—and attest that the model accurately predicts the POAG risk. The presence of background retinopathy was an exclusion criterion in the OHTS. "It is not clear whether these models also predict the progression of established disease or the development of visual disability from the need for glasses, retinopathy, cataracts and other diseases."

2.4 STAR Model

While age and intraocular pressure are reinforcing or triggering factors, the other one to four parameters describe, specify or validate the loss of vision. In this respect, glaucoma risk calculators take 3-6 parameters into account, https://oil.wilmer.jhu.edu/risk/

"The STAR is an easy-to-use slide rule with which physicians enter six risk factors—the five mentioned above, plus diabetes mellitus, which in OHTS appeared to have a protective effect (cf. Fig. 2)." [6]

Age	65	years (40-90)
Baseline IOP	22	mmHg (22-32)
Pattern Standard Deviation	2.0	dB (0.50-4.00)
Central Corneal Thickness	532	microns (450-700)
Vertical Cup to Disc	0.7	0 to 0.9
Diabetes	◉ No	○ Yes
Calculate Risk	Clear Values	
Risk Estimate	40.1	percent in 5 years

This calculator is based on the results described by Medeiros et al. [8]

Fig. 2. The STAR calculator based on the combined analysis of the OHTS and EGPS

The following log-log regression (LLR) formula describes the 5-year risk of developing POAG:

POAG Risk R = $1–0.906^{\exp(PI)}$, where PI:=0.0223·age + 0.1044·IOP + 1.1157·PSD – 0.0150·CCT + 2.7763·VCD – 1.0498·diabetes mellitus + 1.6904

Relevant parameter intervals: Age 40–72, IOP 22–28, VCD 0.2–0.6, CCT 650–500, PSD 1–3. Introducing the mean values for the parameters of an example patient with

age=65 y, IOP= 25, PSD=2, CCT=600, VCD =0.4, no diab melitus (i.e. value 0), we derive PI= 0.09182. For the patient, this results in

POAG Risk =1−exp(ln(0.906)·exp(0.09182)) = 1–0.89744 ~0.1026

ln 0.906 ·exp(PI) = −ln (1−R) ≈ 0.1·exp(PI), R<< 1 → R≈ 0.1·exp(PI)= 0.1096.

R=10.26%; STAR calculator 10.4%.

From now on we will assume the following mean values for a patient cohort:

Age=56, IOP=25, VCD=0.39, CCT=573, PSD =1.9 and the POAG Risk 10.91%, resp. STAR Calculator R=11.1%

Introducing patient's mean values, then the evaluation of the weighted sum in our formula results in -2.654943 and we find

POAG R= $1-0.906^{\exp(PI)}$, where
PI =0.0223 · (age–56) + 0.1044 ·(IOP–25) + 1.1157·(PSD–1.9) – 0.0150·(CCT–573) + 2.7763 · (VCD–0.39) – (1.0498 · diabetes mellitus) + 0.156787

And finally, if we use the fuzzy logic terminology, the crisp values representing weights and parameters could be replaced by intervals to bound individual risk for glaucoma occurrence in family history and ethnicity.

2.5 Three Parameter-Lower Bound of POAG Risk

Based on 5-year risk scores derived from the already introduced Glaucoma risk calculators STAR and WPOAG for progressive glaucoma disease depending on age, IOP and CCT, we develop a DSM with explicit weight functions for the three basic parameters age, IOP, CCT, and furthermore VCD and PSD.

Risk factors age (a), IOP (I) and the weights m_a, m_I as contextual factors influencing CCT, with m_{CCT} as an indicator of visual acuity deterioration. Our DSM-STAR and DSM-WPOAG give lower bounds for the 5-year glaucoma risk and an additional term for individuals with African ancestry. A model patient group uses for calibration the means age:=56, IOP:=25, VCD:=0.39, CCT:=573, and PSD:=1.9 which results in a risk value of 11.1%. for the STAR calculator, 8.9% for the WUSM POAG Risk Calculator, 5 Pts. Low risk for the LGRC, and DSM yields 10.18%.

DSM-STAR: Age: 55-88y, IOP: 22-32, CCT: 450-600, (PSD: 1.9, VCD: 0.39) m_a= (A-56)/3, m_I=4/3(IOP-25), $m_{CCT}= 7 \cdot 2^{(600-CCT)/50}$ [11] $m_{a\cup I}= (m_a + m_I)$ if $m_a, m_I > 0$, $m_{a,eth}= 2+\lfloor(A-55)/7\rfloor$ for special ethnicities. Instead of m_{CCT} you can also use $m_{PSD}=11.1\cdot 2.62^{PSD-1.9}$ or $m_{VCD}= 10.355+ 0.5\cdot((10\cdot VCD-3.9)^2- (10\cdot VDC-3.9)) +3.9\cdot(10\cdot VCD -3.9)$
DSM-WPOAG: $m_{A\cup IOP} -8.9 = 0.05((A-56)/4+IOP-25)^2+ 0.7((A-56)/4+ (IOP-25))$ $m_{PSD\cup VCD}=c_0+c_1(10VCD-3.9)+c_2(10VCD-3.9)^2+d_1(PSD-1.9)+d_2(PSD-1.9)^2$, $d_2=-16$, $d_1=32$, $c_2=0.15$ $c_1=1.5$, $c_0=8.9$; $m_{CCT}=8.9\cdot 2.0^{(573-CCT)/42}$ x= (A–56)/24 + (IOP–25)/6 + (573–CCT)/42; $m_{A\cup IOP \text{ and } CCT} = 1.3\ x^2-0.3x$; y= (573–CCT)/42+(10VCD–3.9)/3.5+(PSD–1.9); $m_{CCT \text{ and } VCD\cup PSD} = 1.6y^2-0.6y$, z= (A–56)/24+(IOP–25)/6+(10VCD–3.9)/3.5+(PSD–1.9); $m_{A\cup IOP \text{ and } VCD\cup PSD}=1.3\ z^2- 0.5z$

Examples: Patient with age:=70, IOP:=26, VCD:=0.59, CCT:=550, and PSD:=2.9. DSM-STAR: R_{cct}= 26%, R_{PSD}= 41.1% (STAR 41.9%; WPOAG 23.2%), R_{VDC}=30,2% (STAR 26.7%; WPOAG 18.3%), R=26+17.98+7.8=51.78% (STAR 73.7%; WPOAG: 43.2%).

Patient 62y, IOP=27, CCT=531, VCD=5.9, PSD=2.2: DSM-WPOAG: $m_{A\cup IOP}$=11.9625, $m_{VCD\cup PSD}$ =14.87, m_{CCT}=17.8, $m_{CCT\ and\ VCD\cup PSD}$ =1.8714, $m_{A\cup IOP\ and\ CCT}$ =2.709, $m_{\ CCT\ and\ VCD\cup PSD}$ = 4.515, $m_{A\cup IOP\ and\ VCD\cup PSD}$ =2.024, $m_{\ A\cup IOP\ \cup VCD\cup PSD}$ = 19.925, $m_{\ A\cup IOP\cup CCT}$ = 23.709, $m_{\ CCT\cup VCD\cup PSD}$=28.285. m(62, 27, 531, 5.9, 2.2) = 19.925 + (23.71 + 28.29)/2–8.9= 37.03 (37.4 WPOAG)

Patient with y, IOP, CCT, VCD=0.39, PSD 1.9; DSM-STAR, STAR, WPOAG risk)

- 56, 25, 600; risk = 7; 7.5 STAR; 5.6 WPOAG
- 57, 26, 575; risk= $10/3 + 7 \cdot 2^{½}$ = 13.23; 12.1 STAR; 12.6 WPOAG
- 65, 22, 532; risk = 3 – 4+17.97 =16.97; 17.6 STAR; 16.7 WPOAG
- 80, 25, 550; risk = 8 + 14 = 22; 24.6 STAR; 21.9 WPOAG
- 60, 27, 500; risk = 4/3 + 8/3 + 12/3 + 28 =36; 37.6 STAR; 36.5 WPOAG
- 80, 30, 500; risk = 2(24/3 + 20/3) + 28 = 57.33; 63.5 STAR; 51.3 WPOAG
- 70, 28, 450; risk = 2(14/3+4)+56=73.33; 74.9 STAR; 62.5 WPOAG
- 80, 30, 450; risk = 2(24/3 + 20/3) + 56 = 85.33; 88.1 STAR; 77.3 WPOAG (475)
- 88, 32, 450; risk = 2(32/3+28/3) + 56 = 96; 95.7 STAR; 83 WPOAG (475)
- 75, 28, 475; risk = $2(19/3 +4) + 7 \cdot 2^{2.5}$ = 60.26; 65.5 STAR; 66.7 WPOAG

Example patient for presented online calculators for 5-year risk POAG:
Age: 55-year-old white
Baseline IOPs for right and left eyes are 22 resp. 26 mm Hg
VCD for right and left eyes are 0.4 and 0.4
CCT measurements are 532 and 548 microns
PSD 2.2 dB in each eye.
The mean of the values for the right and left eyes is averaged for each eye-specific predictor, and the points are summed
Results: LGRC: 6 points High risk; OHTS calculator: 11 points; 5-year risk of developing POAG 20%; STAR model: 21.6%; continuous WUSM POAG OHTS/EGPS Risk Calculator 16.9%; DSM-STAR: 14.42%; 16.42% for black ethnicities.

3 Understanding the Genetics of Glaucoma

As described in [14], glaucoma represents a multifactorial disease resulting from a combination of genetic and environmental factors. In a comprehensive multi-trait glaucoma study, researchers have recently developed a polygenic risk score (PRS) metric that, in addition to known risk factors, can predict earlier age at first diagnosis, faster progression of the disease in the early stages, and the need for glaucoma surgery. A number of genes and polymorphisms–the simultaneous occurrence of two or more discontinuous genotypes or alleles in a population–with different inheritance patterns are cited as the cause. Complex inheritance patterns are responsible for the common adult-onset forms. Mendelian inheritance is typical for rare, early-onset diseases; most common Mendelian forms of primary open-angle glaucoma (POAG) are caused by mutations in the myocilin (MYOC) gene, with a prevalence of 2–4% in adult-onset POAG patients [13].

Genome-wide association studies (GWAS) have contributed to the identification of new disease-related loci. Currently, over 100 genomic regions are known to be associated with POAG susceptibility. Apart from gene variants directly linked to the disease, many genetic loci have been associated with risk factors for POAG as Age, IOP, VCD. etc.

Normal-tension glaucoma (NTG) is an OAG characterized by an IOP within a statistically normal range (≤21 mmHg) and has mostly complex genetic basis. The damage to the optic disc, retinal nerve fiber layer, and visual field is different from the damage seen in POAG. Generally, NTG exhibits a complex genetic foundation. However, the analysis of several pedigrees has revealed that approximately 2% of NTG forms, referred to as Familial NTG, are attributed to a single-gene mutation [13].

In the Rotterdam population-based study, first-degree relatives of glaucoma patients underwent standardized examination, including perimetry, and results demonstrated that the risk of developing glaucoma was 9.2 times increased in relatives of glaucoma patients [14].

3.1 Future Directions

A plethora of GWAS focusing on single nucleotide polymorphisms (SNPs) have identified over a hundred genetic markers associated with glaucoma risk. The challenge lies in identifying the actual effector genes responsible for disease pathogenesis.

New approaches should be developed to perform whole exome-sequencing to evaluate the contribution of protein-changing coding-sequence genetic variants to glaucoma risk. Large-scale population-based studies and subsequent studies in families have the potential to uncover such genetic variants that enhance our understanding of the pathogenesis of glaucoma [15].

As far as we can see, there is not yet a Glaucoma risk calculator that includes PRS in the underlying model.

4 Conclusion

In this work we analyze existing online tools for entering the results of ophthalmology examinations that show users a percent-time span statement about their personal glaucoma risk based on an adequate model with two to five eye health parameters.

Risk modeling and mathematical approaches are employed to arrive at the predictions, which are then compared and critically assessed (cf. Table 1). An essential role is played by databases that store important information on the various forms of diseases, their frequency of occurrence, prevention and treatment. Parameters maintained and made available by (inter)national institutions are

- type and subtype; stages of disease
- appropriate treatments according to the patients
- their origin, age, sex, time of first diagnosis, and results of screening examinations
- comparable disease patterns in the family
- DS-like models with 2 to 5 health parameters giving a lower bound for POAG risk are presented and compared with the known 5-year POAG risk models.

Although discrete score-based methods are easy to use, small changes to the input at the limits of the data intervals sometimes lead to major changes in the percentage risk

or the risk class. In addition, it is not clear how the point result is related to the current and future impairment of visual acuity and the visual field. A point-based model masks the dependencies of the recorded variables and their non-linear behavior.

Finally, an explicit dependence of the risk function on statistical parameters such as the mean and variance of (the variables) year, IOP, etc. and other characteristics of the cohorts and their individuals, their origin and family history should be used for model calibration.

Table 1. Comparison of the Presented Online Glaucoma Risk Calculators

Calculator name	Model	Information to cohorts and validation	No. of model parameter	Risk classes	Risk in %
Laroche	S	Yes/Yes	3	2	Yes
OHTS	S	Yes/Yes	6	3	Yes
WPOAG	C	Yes/Yes	5		Yes
STAR	C	Yes/Yes	6		Yes
DSM	E	Yes/Yes	3-5 (+1)*		Yes

Continuous LLR (C); Score-based (S); Explicit, explainable weights (E) based on STAR/WPOAG; * special ethnicities

Changes and tightening in the definition of risk classes over the last 20 years, advances in screening techniques and reporting systems, gene panel testing and the use of related genomics and biomarkers have had a significant impact on the models and algorithms underlying glaucoma risk calculators.

Ethnically mixed cohorts, missing data on patients, their disease and family history, and different standards of digital examinations and analyses performed may lead to very different results and the non-comparability of risk calculators, mainly due to epistemic uncertainty, while the authors cited in the references mainly focus on aleatory uncertainty [3].

As a conclusion, we would like to summarize some of the advantages or drawbacks and derive some minimum requirements for the Risk Scoring Online tools:

Continuous OHTS calculator. A validated model, no random allocations for results between risk classes, but requires adjustment to individual patient group data and does not reflect epistemic uncertainty. Log-log regression models are difficult to understand and to extend with additional parameters.

DSM. A reengineering approach based on the OHTS-model. A comprehensible dependence of risk classes on model parameters, easily calibrated by using weight functions and additional weights depending on the patient groups considered, their origin and statistical mean values for the life and various heath parameters; only indirect validation via existing risk tools and their evaluation.

Important requirements. Adequate information about their purpose, their operation and the handling of the output results is needed.

- Comprehensible information must be available for each particular question

- What kind of information is expected about the individual's demographics, lifestyle, health status and family disease history?
- Depending on the disease pattern, examination outcomes, and patients' own medical lab samples (e.g., biomarkers), patients are assigned to a risk class that is clearly described
- Data and results must also be at your disposal over a longer period of time
- Questionnaires could be completed in a collaborative manner by patients and their doctors
- Adequate information about the questionnaires, their purpose, their operation and the handling of the output results is needed. Experts and users should provide and receive references to relevant literature on data, models and algorithms, as well as recommendations from national and international health associations and essential standards
- Comprehensible information must be available for each particular question.
- What kind of information is expected about the individual's demographics, lifestyle, health status and family disease history?
- Optionally, the questionnaires should be completed jointly by the patients and their doctors, even in several stages, or alternately if data is missing
- Depending on the disease pattern, examination outcomes, and patients' own medical lab samples (e.g., biomarkers) patients are assigned to a risk class that is clearly described.
- Outcomes should be designed in such a way that users are provided with appropriate counseling and help services depending on their allocation to a risk class, including references to the effects of various sources of uncertainty.
- Data and results must also be at disposal over a longer period of time.
- References to evaluation procedures and metrics that have been carried out make it possible to assess the quality in terms of accuracy, performance and interpretability, involving the various groups involved in the planning, implementation, and use of the service [3].

Acknowledgement. We would like to thank the ophthalmologist Mr. Matthias Brab MB, BAO, BCh (IRL) for giving us much valuable information about glaucoma.

References

1. American Academy of Ophthalmology: Risk Calculators for Primary Open-Angle Glaucoma. Oct. 10, 2019. https://www.aao.org/education/interactive-tool/risk-calculators-primary-open-angle-glaucoma
2. Auer, E., Luther, W.: Uncertainty Handling in Genetic Risk Assessment and Counseling. Journal of Universal Computer Science **27** (12), 1347–1370 (2021)
3. Baloian, N., Luther, W., Peñafiel, S., Zurita, G.: Evaluation of Cancer and Stroke Risk Scoring Online Tools, in Hajian, A. Baloian, N., Inoue, T., Luther, W. (eds.): Data Science, Human-Centered Computing, and Intelligent Technologies. Logos, Berlin, 106–111 (2022)
4. Gordon, M. O., Torri, V., Miglior, S., et al.: A Validated Prediction Model for the Development of Primary Open Angle Glaucoma in Individuals with Ocular Hypertension. Ophthalmology **114**(1), 10–19 (2007)
5. Kass, M. A. et al.: Assessment of Cumulative Incidence and Severity of Primary Open-Angle Glaucoma Among Participants in the Ocular Hypertension Treatment Study After 20 Years of Follow-up. JAMA Ophthalmol. **139**(5), 1–9 (2021)

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8050785/?report=printable
6. Karmel, M.: Glaucoma: Calculating the Risk. EyeNet Magazine (2024) https://www.aao.org/eyenet/article/glaucoma-calculating-risk
7. Laroche, D., Rickford, K., Mike, E. V., Hunter, L., Ede, E., Ng, C., Douglas, J.: A Novel, Low-Cost Glaucoma Calculator to Identify Glaucoma Patients and Stratify Management. Hindawi J. of Ophthalmology, Article ID 5288726, 6 p. (2022)
8. Medeiros, F. M. et al.: Long-term Intraocular Pressure Fluctuations and Risk of Conversion from Ocular Hypertension to Glaucoma. Ophthalmology **115**(6), 934–940 (2008)
9. Risk Calculators for POAG. https://oil.wilmer.jhu.edu/risk/
10. Shaw, J.: OHTS: 20 Years of Follow-up Data on POAG. EyeNet Magazine June 2021 https://www.aao.org/eyenet/article/ohts-20-years-of-follow-up-data-on-poa
11. Siegfried, C. J., Shui, Y. B.: Racial Disparities in Glaucoma: From Epidemiology to Pathophysiology. Mo Med. **119**(1), 49–54 (2022)
12. Tham, Y.C., Li, X., Wong, T.Y., Quigley, H.A., Aung, T., Cheng, C.Y.: Global Prevalence of Glaucoma and Projections of Glaucoma Burden Through 2040: A Systematic Review and Meta-Analysis. Ophthalmology. **121**(11), 2081–90 (2014)
13. Tirendi, S., Domenicotti, C., Bassi, A. M., Vernazza, St.: Genetics and Glaucoma: The State of the Art. Frontiers in Medicine **10**, 11 p. (2023)
14. Topouzis, F. Giannoulis, D.: Understanding the Genetics of Glaucoma. Ophthalmology Times Europe**17**(4) May (2021) https://europe.ophthalmologytimes.com/view/understanding-the-genetics-of-glaucoma
15. Wang, Z., Wiggs, J. L., Aung, T., Khawaja, A. P., Khor, C. C.: The Genetic Basis for Adult-Onset Glaucoma: Recent Advances and Future Directions. Progress in Retinal and Eye Research **90**, Paper ID 101066 (2022)
16. World Health Organization: Blindness and Vision Impairment. Key Facts, 10 August 2023 https://www.who.int/news-room/fact-sheets/detail/blindness-and-visual-impairment
17. Zukerman, R., Harris, A., Vercellin, A. V., Siesky, B., Pasquale, L. R, Ciulla, T. A.: Molecular Genetics of Glaucoma: Subtype and Ethnicity Considerations. Genes (Basel) **12**(1) 55 (2021)

Color Image Enhancement with Quaternion Fourier Transform-Based Alpha-Rooting

Anna A. Vardazaryan[1] and Artyom M. Grigoryan[2][0000−0001−6683−0064]

[1] Yerevan State University, Yerevan, Armenia
anna.vardazaryan3@edu.ysu.am
[2] ECE Dept, The University of Texas at San Antonio, San Antonio, TX 78249, USA
amgrigoryan@utsa.edu
https://ceid.utsa.edu/agrigoryan/

Abstract. This work presents a recent new effective method in color image enhancement method using the traditional non-commutative quaternion arithmetic for color images. In this arithmetic, the RGB color image together with the gray image is presented as a 4-component quaternion image. The method of alpha-rooting with the 2-D quaternion discrete Fourier transform (QDFT) for processing color images is described. Unlike traditional methods that enhance each color component individually, which often leads to color artifacts, the proposed method effectively processes all colors as one unit. The quaternion-based approach preserves the natural relationship among image components. The results of color image enhancement by the proposed method and comparison with the conventional color enhancement methods like color histogram equalization and channel-by-channel enhancement using the 2-D discrete Fourier transform-based alpha-rooting are described. Drone-captured images were utilized to showcase the practical application and effectiveness of our image enhancement method. The findings demonstrate that quaternion-based approach is more effective at preserving the color relationships and features of the image, offering a significant advancement in the field of image enhancement.

Keywords: Image enhancement · Alpha-rooting · Quaternion · Fourier transform.

1 Introduction

In the realm of digital image processing, image enhancement stands out as a crucial technique aimed at augmenting the quality of images for subsequent analysis or improved visual perception. Unlike general image processing, which encompasses a wide array of operations from acquisition to analysis, image enhancement is specifically tailored to modify the appearance of an image in a manner that is more pleasing to the observer or more suitable for analysis. Various factors can degrade image quality, including limitations of the capture device, poor lighting conditions, night scenes, dark objects, bright backgrounds, and other environmental challenges. As a result, essential details within images often

remain obscured, necessitating effective restoration techniques. Image enhancement encompasses a spectrum of techniques tailored to specific application needs, aimed at altering the appearance of images to achieve optimal results. These results include, but are not limited to, contrast enhancement, noise reduction, and color correction. The application of image enhancement is extensive and diverse, encompassing fields such as medical imaging, underwater imaging, biometrics, aerial and satellite images, thermal images, pattern recognition, remote sensing, and computation photography.

The techniques used for image enhancement are varied and tailored to address specific challenges; no single technique is universally effective. They are categorized into two primary groups: spatial domain methods and frequency domain methods. Spatial domain methods involve the direct manipulation of pixel values within an image (such as Histogram equalization [1][3], Retinex methods [4], and Adaptive gamma correction [5]). On the other hand, frequency domain methods transform the image into frequency space, facilitating the manipulation of specific frequency components (such as the Fourier, Hartley, Hadamard, cosine and other transforms [6]). We focus on the effective method of α-rooting [7] based on Fourier transform as experimental results show that the most effective transform for alpha-rooting is the Fourier transform.

The motivation of this paper is to improve the clarity and vibrancy of color images captured by drones, which are increasingly vital in fields such as agriculture, disaster response, and military.

2 Methodology of 2D DFT and 2D QDFT Based Alpha-Rooting

2.1 2-D Discrete Fourier Transform

The two-dimensional discrete Fourier transform (2-D DFT) of the grayscale image $f_{n,m}$ at the frequency-point (p, s) is defined as [2]:

$$F_{p,s} = \sum_{n=0}^{N-1} \sum_{m=0}^{M-1} f_{n,m} W_M^{ms} W_N^{np} = \sum_{n=0}^{N-1} \left[\sum_{m=0}^{M-1} f_{n,m} W_M^{ms} \right] W_N^{np},$$

where, $p = 0 : N-1$ and $s = 0 : M-1$. The transform $\{F_{p,s}\}$ is called the $N \times M$-point 2-D DFT of the image $f_{n,m}$. The basis functions are defined by the exponential coefficients: $W_K = e^{-2\pi i/K} = \cos(2\pi/K) - i \cdot \sin(2\pi/K)$, for $K = N, M$.

The result of a 2D DFT on an image is a complex matrix where each element represents a particular frequency component. These complex numbers have both magnitude $|F_{p,s}|$ and phase $\psi(p, s) = \text{phase}[F(p, s)]$, $p = 0 : (N-1)$, $s = 0 : (M-1)$. Here magnitude tells us how strong a particular frequency is in the image (often corresponds to edges and other abrupt changes in intensity) and phase indicates the position of these frequency components within the image.

2.2 Quaternion Algebra

Quaternions were first discovered by by the Irish mathematician Hamilton in 1843. They are a 4-dimensional generalization of complex numbers and are defined as follows [2][10]:

$$q = a + bi + cj + dk,$$

where a, b, c, and d are real numbers and the units have the following properties:

$$i^2 = j^2 = k^2 = ijk = -1,$$
$$ij = -ji = k, jk = -kj = i, ki = -ik = j.$$

Thus, the quaternion q is presented in the form $q = a + (bi + cj + dk) = a + q'$, where a is real and q' is three-component imaginary part.

The magnitude of the quaternion is defined as follows:

$$|q| := \sqrt{\|q\|} = \sqrt{a^2 + b^2 + c^2 + d^2}.$$

When norm of the quaternion $\|q\|$ is equal to 1, q is called a unit quaternion, and when the real part is equal to 0, it is called a pure quaternion.

The addition and multiplication of quaternions follow associative rules similar to those in traditional algebra. However, the multiplication of quaternions is not commutative, a consequence of the specific product rules of their basic components. That is, for the given quaternions p and q: $pq = qp$ or $pq \neq qp$.

Color Image as a Quaternion Number Color images can be expressed within quaternion space, which typically involves three or four channels depending on the color model used. For three-channel systems such as RGB or XYZ, it's possible to describe color images as pure quaternions. For instance, in the RGB model, the quaternion representation would be [9]:

$$q(n, m) = ir(n, m) + jg(n, m) + kb(n, m),$$

where $r(n, m)$, $g(n, m)$ and $b(n, m)$ represent the red, green, and blue channels, respectively. The brightness as the real part can also be added.

This quaternion approach to representing color images creates an integrated relationship among all the color components, which is something that conventional color enhancement methods, that process channels separately, do not achieve.

2.3 2-D Discrete Quaternion Fourier Transform

As the quaternion multiplication is not commutative, the definition of a 2-D Discrete Quaternion Fourier Transform is not unique. Depending on the position of the transformation kernel relative to the signal, the following transformations are defined: the two-side 2-D QDFT, the right-side and left-side 2-D QDFTs [2][12].

We consider the two-side 2-D QDFT, which is defined for a quaternion image $q_{n,m}$ of size $N \times M$ as follows:

$$Q_{p,s} = \sum_{n=0}^{N-1} \sum_{m=0}^{M-1} W_j^{np} q_{n,m} W_k^{ms} = \sum_{n=0}^{N-1} W_j^{np} \left(\sum_{m=0}^{M-1} q_{n,m} W_k^{ms} \right),$$

where $p = 0 : N-1$ and $s = 0 : M-1$. The components $F_{p,s}$ of the transform are quaternion numbers with real and imaginary parts.

The inverse 2-D two-side QDFT is calculated by the similar formula:

$$q_{n,m} = \frac{1}{NM} \sum_{p=0}^{N-1} \sum_{s=0}^{M-1} W_j^{-np} Q_{p,s} W_k^{-ms} = \frac{1}{NM} \sum_{p=0}^{N-1} W_j^{-np} \left(\sum_{s=0}^{M-1} Q_{p,s} W_k^{-ms} \right),$$

where $n = 0 : N-1$ and $m = 0 : M-1$.
The basis functions are defined by the exponential coefficients:

$$W_j = \cos(2\pi/N) - j \sin(2\pi/N) \quad \text{and} \quad W_k = \cos(2\pi/M) - k \sin(2\pi/M).$$

2.4 Image Enhancement Measures

To measure image quality and select optimal processing parameters, we consider enhancement metrics based on Weber's law of the human visual system [2][11]. The image is divided by blocks, let say 7x7, and max and min are calculated in each block.

The Enhancement Measure Estimation(EME) of an $N \times M$ sized $f_{n,m}$ grayscale image is based on the concept of contrast perception and is defined as

$$EME(f) = \frac{1}{k_1 k_2} \sum_{k=1}^{k_1} \sum_{l=1}^{k_2} 20 \ln \left(\frac{\max(f_{k,l})}{\min(f_{k,l})} \right).$$

For the RGB color model, an analogous metric is established as

$$CEME(f) = \frac{1}{k_1 k_2} \sum_{k=1}^{k_1} \sum_{l=1}^{k_2} 20 \ln \left(\frac{\max_{k,l}\{f_R, f_G, f_B\}}{\min_{k,l}\{f_R, f_G, f_B\}} \right),$$

where f_R, f_G, f_B are the red, green and blue channels of the image, respectively.

2.5 Fourier Transform-Based Alpha-Rooting Method

The α-rooting image enhancement technique can be used to enhance edge information and sharp features in images, as well as for enhancing even low contrast images [7][8]. In the alpha-rooting method of image enhancement for each frequency point (p, s), the magnitude of the discrete Fourier transform is transformed as $|F_{p,s}| \to |F_{p,s}|^{\alpha}$. In the case of the quaternion approach, the process remains consistent, the primary difference is that we apply the Quaternion Discrete Fourier Transform (QDFT) in place of the standard DFT.

Here, the parameter α is from the range $(0, 1]$ and in practice it is generally chosen from a smaller range, for example $[0.4, 1]$.

Diagram of the presented method of image enhancement is given in Fig. 1.

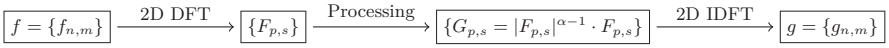


Fig. 1. Block-diagram with the Fourier transform-based alpha-rooting enhancement

As an example, Fig. 2 shows the original grayscale image "Ararat1.jpg" in part (a), the graph of the EME in part (b), and results of the α-rooting for α=0.66, 0.87, and 0.85 in parts (c),(d), and (e), respectively.

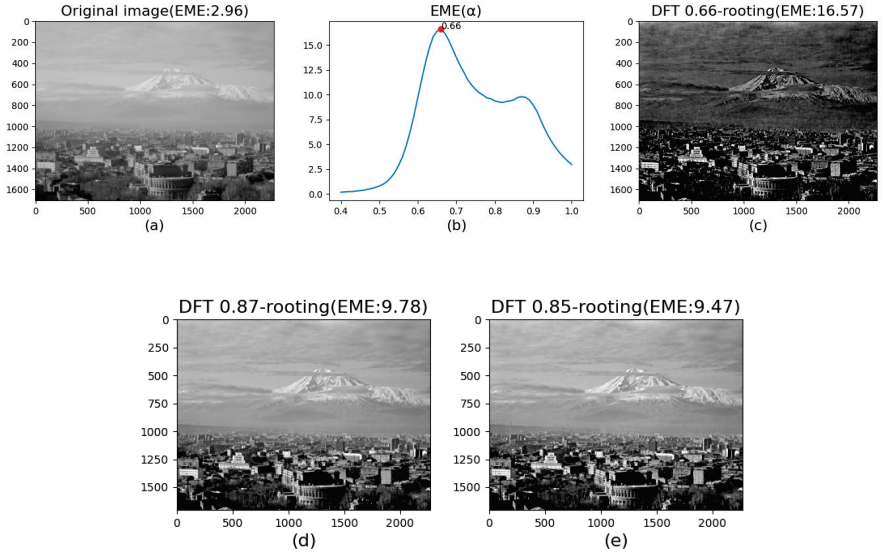
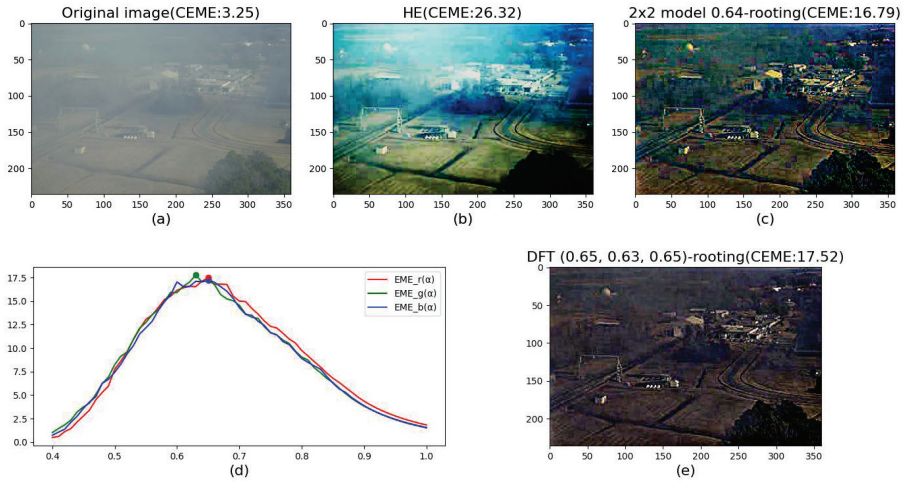


Fig. 2. (a) Original "Ararat1.jpg" image (b) $EME(\alpha)$ graph (c) DFT 0.66-rooting (d) DFT 0.87-rooting (e) DFT 0.85-rooting image enhancement

3 Experimental Results



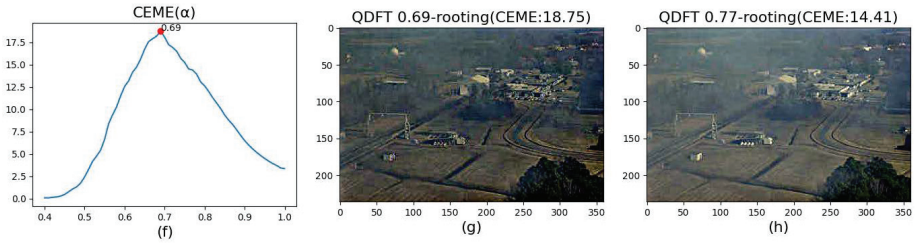


Fig. 3. (a) Original "field.jpg" image (b) HE of the image (d) $EME(\alpha)$ and (f) $CEME(\alpha)$ graph (c) 2×2 grayscale model DFT 0.66-rooting (e) DFT (0.65, 0.63, 0.65)-rooting (g) QDFT 0.69-rooting (h) QDFT 0.77-rooting image enhancement

Figure 3 presents a comparative analysis of image enhancement techniques applied to an original image of poor quality, shown in part (a). Histogram Equalization (HE), applied to enhance contrast, demonstrates an increase in CEME, as illustrated in part (b). This technique adjusts the dynamic range to better utilize available tones. A 2×2 grayscale model [2] DFT-based alpha-rooting method is then employed, optimizing the alpha parameter based on the highest EME value observed. This approach, shown in part (c), enhances finer details by modulating the frequency components. Further, the image is processed using a channel-by-channel 2-D DFT-based alpha-rooting method, as seen in part (e). Here, the optimal alpha values for each color channel (red, green, and blue) are determined based on their respective EME vs. alpha plots, ensuring that detail clarity is maximized for each channel, as depicted in the plot in part (d). Part (g) displays the superior results of QDFT-based enhancement, where the chosen alpha value is derived from the CEME vs. alpha graph presented in part (f). Additionally, an alternative alpha setting is tested, with its results shown in part (h), allowing for a comparative analysis to assess how varying alpha levels influence the overall image quality. The values of the EME and CEME for the processed above images are given in Table 1.

Table 1. Alpha and EME/CEME values of original and enhanced images

"field.jpg"	**CEME**	**Alpha**	**EME**
Original Image	3.25		R: 1.82 G: 1.57 B: 1.52
Histogram Equalization	26.32		
2-D DFT Alpha-Rooting	17.52	R: 0.65 G: 0.63 B: 0.65	17.48 17.76 17.21
2-D DFT Alpha-Rooting of 2-D Grayscale Image(2x2 model)	16.79	0.64	
2-D QDFT Alpha-Rooting	18.75	0.69	
2-D QDFT Alpha-Rooting	14.41	0.77	

One can note that HE image is very bright on the top part of it and has false blue colors. This is why its CEME is a large number, 26.32. All other images with alpha-rooting have high enhancement measure.

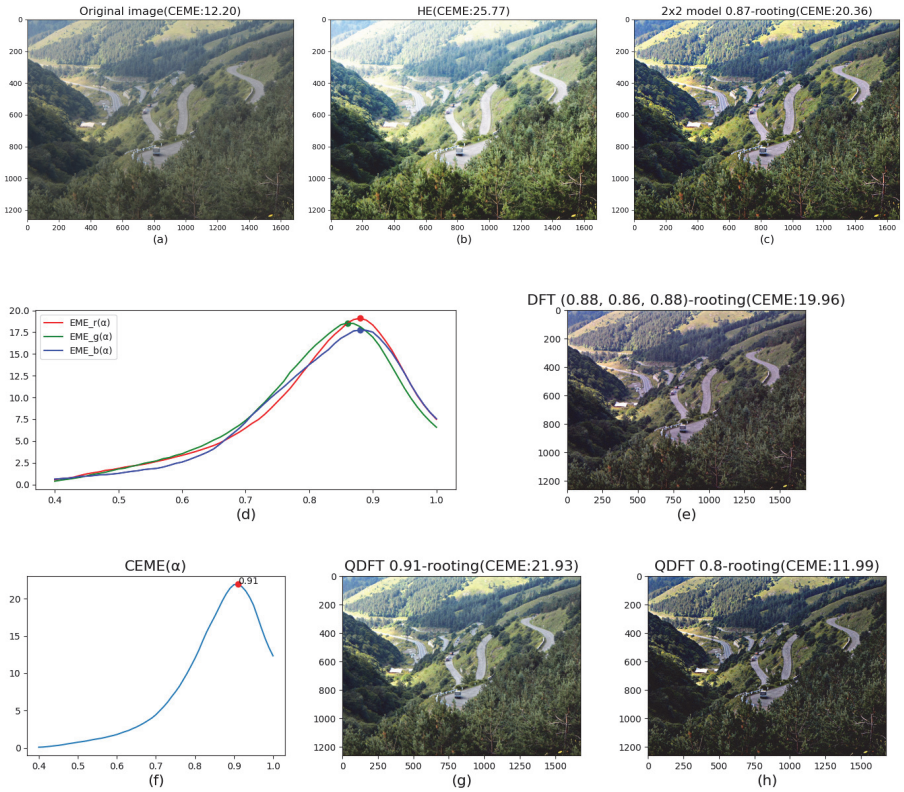


Fig. 4. (a) Original "dilijan.jpg" image (b) HE of the image (d) $EME(\alpha)$ and (f) $CEME(\alpha)$ graph (c) 2×2 grayscale model DFT 0.87-rooting (e) DFT (0.88, 0.86, 0.88)-rooting (g) QDFT 0.91-rooting (h) QDFT 0.8-rooting image enhancement

Fig. 4. replicates the analysis presented in Fig. 3., applying the same enhancement techniques to the "dilijan.jpg" image, with adjustments to the parameter values to accommodate the unique characteristics of this new image. The alpha and EME/CEME values for the processed images are provided in Table 2.

Table 2. Alpha and EME/CEME values of original and enhanced images

"dilijan.jpg"	**CEME**	**Alpha**	**EME**
Original Image	12.20		R: 7.49 G: 6.58 B: 7.47
Histogram Equalization	25.77		
2-D DFT Alpha-Rooting	19.96	R: 0.88 G: 0.86 B: 0.88	19.10 18.55 17.74
2-D DFT Alpha-Rooting of 2-D Grayscale Image(2x2 model)	20.36	0.87	
2-D QDFT Alpha-Rooting	21.93	0.91	
2-D QDFT Alpha-Rooting	11.99	0.80	

All experiments were conducted using Python, employing fast algorithms for the Discrete Fourier Transform [13] and a block size of 7x7 for the EME/CEME calculations.

4 Conclusion

The experimental results demonstrate that the alpha-rooting method, based on the Fourier transform, improves image quality more effectively than histogram equalization. It is important to note that in the case of color images, enhancing channels individually does not take into account the relationship between the colors, leading to artificial colors in the enhanced image. Instead, the proposed quaternion approach enhances images by preserving the original natural colors. The values of the Color image enhancement measure (CEME) corroborate this finding. Compared to the channel-by-channel and 2×2 grayscale model (which takes more than twice as long as the quaternion approach) enhancements, the CEME values are higher for the quaternion discrete Fourier transform.

References

1. Gonzalez, R.C., Woods, R.E.: Digital Image Processing, 3rd edn. Pearson Education, Inc., Upper Saddle River, New Jersey 07458 (2008)
2. Grigoryan, A.M., Agaian, S.S.: Quaternion and Octonion Color Image Processing with MATLAB, PM279, SPIE, Bellingham, WA 98225, USA (2018)
3. Grigoryan A.M., Agaian, S.S.: Novel method of color histogram equalization: Binding the colors with brightness. IJSER **10**(12), 1100-1106 (2019)
4. Land, E.H.: The retinex theory of color vision. Scientific American **237**(6), 108-129 (1977)
5. Gang, C. et al.: Contrast enhancement of brightness-distorted images by improved adaptive gamma correction. Computers & Electrical Engineering **66**, 569-582 (2018)
6. Aghagolzadeh S., Ersoy O.K.: Transform image enhancement. Optical Engineering **31**(3), 614-626 (1992)
7. McClellan, J.: Artifacts in alpha-rooting of images. In: ICASSP'80. IEEE International Conference on Acoustics, Speech, and Signal Processing, Vol. 5, pp. 449-452. IEEE, New York, USA (1980)
8. Grigoryan, A.M., Agaian, S.S.: Image Processing Contrast Enhancement. Wiley Encyclopedia of Electrical and Electronics Engineering, p. 22 (2017)
9. Sangwine, S.J.: Fourier transforms of colour images using quaternion or hypercomplex numbers. Electronics Letters, **32**(21), 1979-1980 (1996)
10. Hamilton, W.R.: Elements of Quaternions. Logmans, Green and Co., London (1866)
11. Agaian, S.S., Panetta, K., Grigoryan, A.M.: A new measure of image enhancement. In: IASTED International Conference on Signal Processing and Communication, pp. 19-22, Citeseer (2000)
12. Yeh, M.H.: Relationships among various 2-D quaternion Fourier transforms. IEEE Signal Processing Letters 15, 669-672 (2008)
13. Cooley, J.W., Tukey, J.W.: An algorithm for the machine calculation of complex Fourier series. Mathematics of Computation **19**(90), 297-301 (1965)

Novel Gradient-Based Retinex Method for Image Enhancement

Armine A. Bayramyan[1] and Artyom M. Grigoryan[2][0000−0001−6683−0064]

[1] Yerevan State University, Yerevan, Armenia
armine.bayramyan@edu.ysu.am
[2] ECE Dept, The University of Texas at San Antonio, San Antonio, TX 78249, USA
amgrigoryan@utsa.edu
https://ceid.utsa.edu/agrigoryan/

Abstract. The retinex method, which includes single and multi-scale algorithms, is an effective method for grayscale and color image enhancement that proved color constantly and dynamic range compression. There are different implementations of the retinex algorithm, which allow different degrees of user control over parameters, color correction, intermediate steps and filters, and different forms of application. This study introduces a novel enhancement technique known as gradient-based retinex (GB-Retinex), which is uniquely applied to overcome common imaging challenges such as inconsistent lighting and unclear details. The presented method utilizes the Retinex theory, applying it to the low-pass filtered images by means of gradient operators (e.g. symmetric Laplacian gradients) to highlight critical features and enhance contrast. Through comparative analysis with the traditional gradient-based histogram equalization technique, GB-Retinex has shown to provide superior image improvements (we also use the Enhancement Measure Estimation (EME) for the comparison of methods). The enhanced clarity and detail achieved with GB-Retinex make it a preferable choice for image enhancement across different contexts. This advancement not only betters the visual quality but also the functional utility of images for subsequent analysis tasks. Illustrative examples with different images are given together with enhancement by method GB-HE.

Keywords: Image enhancement · Histogram equalization · Retinex · Gradients.

1 Introduction

Image enhancement is a crucial technique in digital image processing that aims to better the visual quality of images for human perception and more efficient computer analysis [1],[2]. It involves a range of methods, such as adjusting brightness, color correction, noise reduction, and sharpening details. These techniques are essential across various sectors; drones use enhanced images for clearer aerial views in agriculture and disaster relief, while thermal imaging is vital in medicine, building inspections, and security for heat detection. Enhanced medical images like X-rays and MRIs facilitate better diagnosis and treatment [3].

In consumer electronics, enhancement algorithms make photos and videos taken by smartphones and cameras more appealing. In security, these techniques help in identifying important features for crime prevention and surveillance. Image enhancement is not one-size-fits-all; it must be tailored to each image and its application. For example, enhancing a medical image to spot fractures is different from enhancing an aerial photo for land use analysis [3].

Current research includes creating algorithms that replicate human vision, methods that maintain color fidelity, and enhancing images for the visually impaired. Enhancements can be in the spatial domain, where the image is processed in its raw form, or the frequency domain, involving data manipulation post frequency conversion [4].

Notably, in this study, we present a new enhancement method and utilize drone imagery to showcase the potential of GB-Retinex in real-world applications. Drone-captured images often present unique challenges due to varying altitudes, angles, and lighting conditions.

2 Retinex for Grayscale Images

The retinex theory, introduced by Edwin Land [5], explains how humans perceive consistent object colors under varying lighting.

- It compresses the dynamic range, meaning it can render a large input dynamic range into a relatively small output dynamic range.
- Sharpens the image, counteracting the blurring that occurs when receiving pictures. This allows small details to be seen more easily than before.
- Provides color constancy, removes the effects of the illumination from the subject.

Retinex uses a multiplicative approach to enhance images. It breaks down an image into its reflectance and illumination:

$$f_{n,m} = l_{n,m} r_{n,m}, \quad (n,m) \in \mathbb{R}^2.$$

Retinex algorithms initiate by converting the original image into the logarithmic domain. This transformation is beneficial because it leverages the product rule of logarithms to discern the difference between the image and its illumination [6].

There are two main types of the retinex algorithm: single-scale retinex (SSR) and multi-scale retinex (MSR).

2.1 Single-Scale Retinex (SSR)

$$J(n,m;\sigma) = \ln\left(\frac{f_{n,m}}{F(n,m) * f_{n,m}}\right) = \ln[f_{n,m}] - \ln[F(n,m) * f_{n,m}],$$

where $f_{n,m}$ is the input image and $F(n,m)$ is the Gaussian function: $F(n,m) = A \cdot \exp\left(-\frac{n^2+m^2}{2\sigma^2}\right)$, σ is a constant which controls the extent of F,

$A = \left(\sum_{n=0}^{N-1} e^{-\frac{n^2}{2\sigma^2}} \right)^{-1} \left(\sum_{m=0}^{M-1} e^{-\frac{m^2}{2\sigma^2}} \right)^{-1}$ and * represents spatial 2-D convolution [7]-[9].

2.2 Multi-Scale Retinex (MSR)

The single-scale retinex sometimes does not provide a high or desired quality of enhancement. For that, we can take the linear combination of a few different SSRs, giving each of them a weight [10],[11]:

$$f_{n,m} = \omega_1 J(n, m; \sigma_1) + \omega_2 J(n, m; \sigma_2) + \cdots + \omega_l J(n, m; \sigma_l).$$

Here l is the number of scales σ_k of the Gaussian filters ($k = 1 : l$),

$$\omega_1 + \omega_2 + \cdots + \omega_l = 1.$$

3 Retinex for Color Images

Color models, sometimes called color spaces, help us define and work with colors in a standardized way. Essentially, a color model is a specification of a coordinate system and a subspace of that system, where each color is represented by a single point [3].

3.1 Retinex in HSV Color Space

The use of retinex methods for color images depends on the color model. In the HSV color model the retinex method is applied only to the V color channel, and the other two components (H, S) remain unchanged:

$$J(n, m; \sigma) = (h_{n,m}, s_{n,m}, J_V(n, m; \sigma)).$$

3.2 Retinex in RGB Color Space

An image in the RGB color space involves processing the red $r_{n,m}$, green $g_{n,m}$, and blue $b_{n,m}$ channels of the image $f_{n,m}$ separately. The enhanced channels are then combined to form a new, improved image.

Employing the single-scale retinex theory, we can manipulate the individual color channels with the same scale factor $s = 2\sigma^2$:

$$r_{n,m} \to J_R(n, m; \sigma), \quad g_{n,m} \to J_G(n, m; \sigma), \quad b_{n,m} \to J_B(n, m; \sigma).$$

The output image is parameterized by σ and is given by:

$$J(n, m; \sigma) = (J_R(n, m; \sigma), J_G(n, m; \sigma), J_B(n, m; \sigma)) .$$

The enhanced formula for the MSR, based on the scale parameter set σ_k, for $k \in \{1, 2, \ldots, l\}$, processes each color channel K (where $K \in \{R, G, B\}$) as follows:

$$(f_K)_{n,m} \to \sum_{k=1}^{l} \omega_k \log \left[\frac{(f_K)_{n,m}}{y_{\sigma_k}(n,m) * (f_K)_{n,m}} \right] = \sum_{k=1}^{l} \omega_k \log (f_K)_{n,m} - \\ - \sum_{k=1}^{l} \omega_k \log [y_{\sigma_k}(n,m) * (f_K)_{n,m}].$$

4 Gradient-Based Histogram Equalization for Grayscale Images

This method is a fairly simple and fast method that, while preserving the range and average intensity of the image, reduces the standard deviation and significantly corrects the histogram graph [12].

Consider a gradient operator, for instance, one of the symmetric Laplacian gradients with the 3×3 matrices:

$$[G] = \frac{1}{4} \begin{bmatrix} 1 & 0 & 1 \\ 0 & -4 & 0 \\ 1 & 0 & 1 \end{bmatrix}, \quad \frac{1}{4} \begin{bmatrix} 0 & 1 & 0 \\ 1 & -4 & 1 \\ 0 & 1 & 0 \end{bmatrix}, \quad \frac{1}{8} \begin{bmatrix} 1 & 1 & 1 \\ 1 & -8 & 1 \\ 1 & 1 & 1 \end{bmatrix},$$

and parameter $\alpha \in (0, 1]$.

The algorithm of the parameterized GB-HE:

1. Calculate the gradient image: $X_g = X * G$.
2. Calculate the difference image: $X_s = X - X_g$.
3. Calculate the histogram equalization of image: $X'_s = HE(X_s)$.
4. Calculate the new image: $Y = \alpha X'_s + X_g$.

The result Y is the enhanced image.

In the 3rd step instead of the traditional HE, other methods of image enhancement can also be used. We will use the retinex method.

5 Gradient-Based Retinex Method

The algorithm is the same as that of the GB-HE, only instead of the HE we use the retinex method.

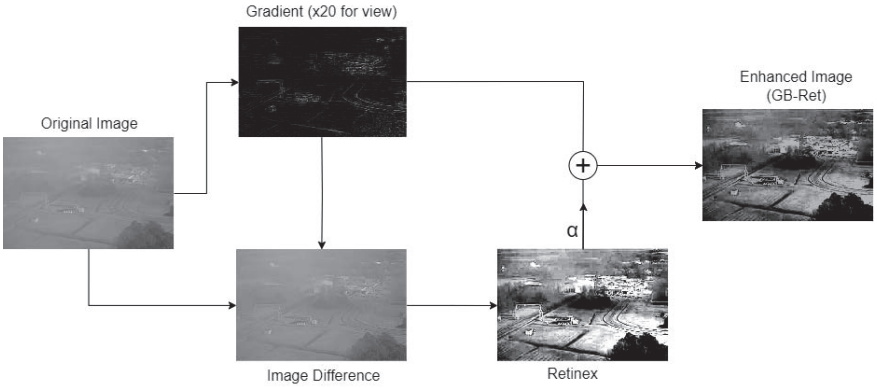


Fig. 1: The block-diagram of the GB-Ret method

Now we compare the results of GB-HE and GB-Ret methods (the gradient

$$[G] = \frac{1}{4}\begin{bmatrix} 1 & 0 & 1 \\ 0 & -4 & 0 \\ 1 & 0 & 1 \end{bmatrix}$$

is considered). The results of processing two images are shown in Figs. 2 and 3.

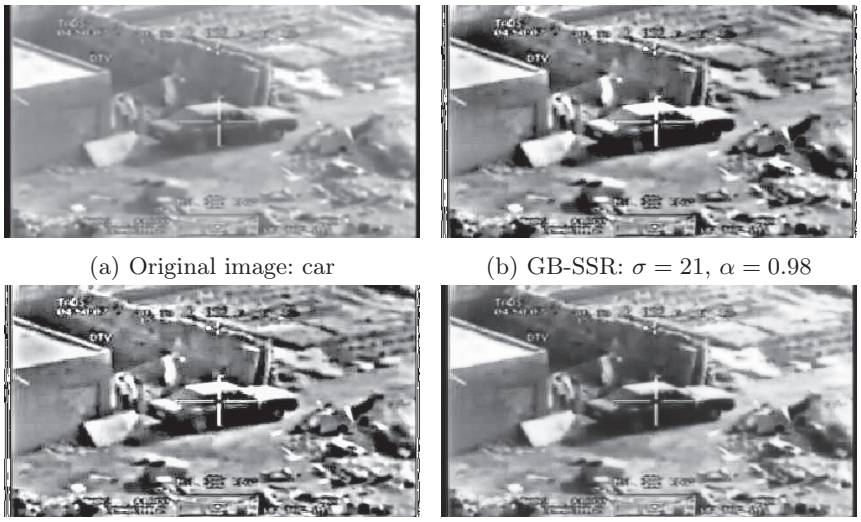
(a) Original image: field

(b) GB-SSR: $\sigma = 45$, $\alpha = 0.7$

(c) GB-MSR: $\sigma = \{15, 31, 45\}$, $\alpha = 0.7$

(d) GB-HE: $\alpha = 0.9$

Fig. 2: (a) Original image, (b) GB-SSR enhanced image, (c) GB-MSR enhanced image and (d) GB-HE enhanced image



(a) Original image: car (b) GB-SSR: $\sigma = 21$, $\alpha = 0.98$

(c) GB-MSR: $\sigma = \{7, 15, 21\}$, $\alpha = 0.98$ (d) GB-HE: $\alpha = 0.98$

Fig. 3: (a) Original image, (b) GB-SSR enhanced image, (c) GB-MSR enhanced image and (d) GB-HE enhanced image

We consider the EME measure which indicates the level of image enhancement by an image processing technique, particularly in terms of image contrast [12]. The image $f_{n,m}$ of size $N_1 \times N_2$ is divided into blocks of size $L_1 \times L_2$ each. Here $k_i = \left\lfloor \frac{N_i}{L_i} \right\rfloor, i = 1, 2$ and $\lfloor \cdot \rfloor$ is the rounding floor function.

$$EME(f) = \frac{1}{k_1 k_2} \sum_{k=1}^{k_1} \sum_{l=1}^{k_2} 20 \log_{10} \left(\frac{\max_{k,l}(f)}{\min_{k,l}(f)} \right).$$

$\max_{k,l}(f)$ and $\min_{k,l}(f)$ respectively are the maximum and minimum of the image $f_{n,m}$ inside the (k, l)-th block.

The EME values of the corresponding images are given in Table 1:

Table 1: EME values of the original and enhanced images (block size is $(8, 8)$)

Image	EME (Original)	EME(GB-SSR)	EME(GB-MSR)	EME(GB-HE)
field	1.61	26.94	28.63	16.54
car	11.39	29.15	31.22	28.46

We can see from this table and also visually from the images that the gradient-based retinex method provides higher quality images than the gradient-based histogram equalization method.

Now, we present two examples for enhanced color images in HSV color space, which are given in Fig. 4 (again the gradient

$$[G] = \frac{1}{4}\begin{bmatrix} 1 & 0 & 1 \\ 0 & -4 & 0 \\ 1 & 0 & 1 \end{bmatrix}$$

is used):

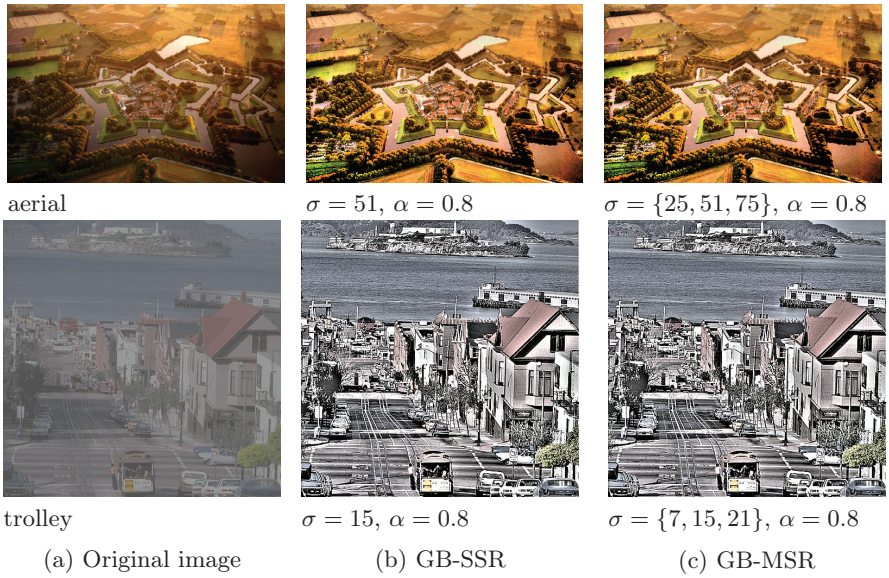


Fig. 4: (a) Original image, (b) GB-SSR enhanced image and (c) GB-MSR enhanced image

We will use the CEME (color image enhancement measure) which is the generalization of the EME measure [13]:

$$CEME(f) = \frac{1}{k_1 k_2}\sum_{k=1}^{k_1}\sum_{l=1}^{k_2} 20\log_{10}\left(\frac{\max_{k,l}(f_R, f_G, f_B)}{\min_{k,l}(f_R, f_G, f_B)}\right),$$

where $\max_{k,l}(f)$ and $\min_{k,l}(f)$ are respectively the maximum and minimum of the image $f_{n,m}$ in the block (k, l). The values of CEME for the above images are given in Table 2.

Table 2: CEME values of the original and enhanced images (block size is $(8, 8)$)

image	CEME (original)	CEME (GB-SSR)	CEME (GB-MSR)
aerial	24.77	26.37	26.60
trolley	6.36	42.44	43.29

Conclusion

This research presents a new image enhancement method, the gradient-based retinex method, for enhancing drone images. The preliminary experimental examples show that the applied technique significantly improves the visual quality of both grayscale and color images. Comparative analysis with gradient-based histogram equalization technique provides convincing evidence that our method outperforms existing approaches, providing superior visual and quantitative benchmark results in drone image enhancement.

Thus, the gradient-based retinex method with its high performance is a promising candidate for further research and application in the field of image processing.

References

1. Plataniotis, K. and Venetsanopoulos, A.N.: Color image processing and applications. Springer Science & Business Media (2000).
2. Wang, D.C., Vagnucci, A.H. and Li, C.C.: Digital image enhancement: a survey. Computer vision, graphics, and image processing **24**(3), pp.363-381 (1983).
3. Gonzalez, C.R. and Woods, E.R.: Digital Image Processing, 3rd edition. Pearson Education, Inc., Upper Saddle River, New Jersey 07458 (2008).
4. Agaian, S.S., Sridharan, V. and Blanton Jr, M.: Switching system for image enhancement and analysis of fused thermal and RGBD data. In Mobile Multimedia/Image Processing, Security, and Applications 2012 (Vol. 8406, pp. 279-291). SPIE (2012, May).
5. Land, E.H. and McCann, J.J.: Lightness and retinex theory. Josa **61**(1), pp.1-11 (1971).
6. Rahman, Z.U., Jobson, D.J. and Woodell, G.A.: Retinex processing for automatic image enhancement. Journal of Electronic imaging **13**(1), pp.100-110 (2004).
7. Jobson, D.J., Rahman, Z.U. and Woodell, G.A.: Properties and performance of a center/surround retinex. IEEE transactions on image processing **6**(3), pp.451-462 (1997).
8. Rahman, Z.U.: Properties of a center/surround Retinex: Part 1: Signal processing design. NASA Contractor Report, 198194, p.13 (1995).
9. Jobson, D.J. and Woodell, G.A.: Properties of a center/surround Retinex: Part 2. Surround design. NASA Technical Memorandum, 110188, p.15 (1995).
10. Rahman, Z.U., Jobson, D.J. and Woodell, G.A.: Multiscale retinex for color rendition and dynamic range compression. In Applications of Digital Image Processing XIX (Vol. 2847, pp. 183-191). SPIE (1996, November).
11. Rahman, Z.U., Jobson, D.J. and Woodell, G.A.: Multi-scale retinex for color image enhancement. In Proceedings of 3rd IEEE international conference on image processing (Vol. 3, pp. 1003-1006). IEEE (1996, September).
12. Grigoryan, A.M. and Agaian, S.S.: Gradient based histogram equalization in grayscale image enhancement. In Mobile Multimedia/Image Processing, Security, and Applications 2019 (Vol. 10993, pp. 132-142). SPIE (2019, May).
13. Grigoryan, A.M. and Agaian, S.S.: Alpha-rooting method of color image enhancement by discrete quaternion Fourier transform. In Image Processing: Algorithms and Systems XII (Vol. 9019, pp. 23-34). SPIE (2014, February).

Fairness in the Use of Medical Online Tools

Wolfram Luther [0000-0002-1245-7628] and Ashot Harutyunyan[0000-0003-2707-1039]
University of Duisburg-Essen, Department of Computer Science, Germany
Yerevan State University, Machine Learning Lab, and
Institute for Informatics and Automation Problems NAS RA, Armenia
wolfram.luther@uni-due.de harutyunyan.ashot@ysu.am

Abstract. The concept of fairness in the development and use of medical risk assessment tools is presented in this paper. After considering various approaches to a general definition of algorithmic fairness from the perspective of the implied sciences, guidelines for system requirements are formulated to highlight the different forms of fairness and their biases. These are for example poor data quality, inadequate models, bad accuracy and performance of algorithms or insufficient interaction or collaboration of stakeholders. The requirements are illustrated using the example of numerous tools for estimating the 5-year risk that an individual with ocular hypertension will develop Primary Open Angle Glaucoma (POAG), the leading global cause of irreversible blindness.

Keywords: Artificial Intelligence, Bias, Algorithmic Fairness, Discrimination, Risk Prevention, Medical Tool, Explainable Artificial Intelligence.

1 Introduction

Artificial intelligence (AI) is a collective term for technologies that support and enhance human abilities in hearing and seeing, analyzing, deciding, communicating and acting. It is based on the comprehensive digitalization of systems and processes (cognitive systems and their digital twins) and requires complex computer-based system architectures and their interfaces. It makes sense to differentiate between partial, complete and extended AI systems and AI-based applications, depending on how extensively human properties of learning, thinking, reasoning and further communication with native languages, facial expressions and gestures between humans and AI systems, and cooperation are enabled and comprehensively supported. This requires large language models and currently huge distributed computing resources.

The tools and technologies used are diverse and depend on the areas in which the AI solutions are running. It is therefore not surprising that special requirements are placed on the results of the use of AI technologies. AI should be explainable, comprehensible and meet specified quality standards [13]. This requires international agreements on domain-based criteria and metrics, procedures for validating results and, if necessary, adapting AI systems through learning in cooperation with experts. Results, predictions, recommendations, and the general use of AI-based assistance systems in the medical sector in particular require interdisciplinary evaluation, assessment and cooperative decision-making when it comes to the specific treatment of patients, preventive or follow-up care. As is known from game theory or multi-objective optimization, it cannot be excluded that not all criteria can be met equally well and that there should be procedures for finding compromises or defining a balance [13].

Of course, all these modern development processes are characterized by economic interests, the systems use existing knowledge, the companies involved use marketable

solutions, data, and metadata that are subject to digital rights management. In that respect, these processes are not conflict-free, not fair and moderation or mediation is required in the event of conflicting interpretations of the results of AI systems.

People who are disadvantaged in terms of access to hardware and software, devices and the tools themselves and people with cognitive impairments can also be the subject of discrimination. In the following, we will look at the concept of (algorithmic) fairness from the perspective of different scientific disciplines.

To identify relevant literature in the field, we conducted a document search in various international databases using the keywords 'explainable', 'artificial intelligence', 'algorithmic fairness', 'bias', 'discrimination', 'health tool', and 'risk prevention' which returned 76 hits in the categories journal articles/reviews, conference papers and (chapters in) books published between 2019 and 2024 from the fields of computer and social sciences and engineering. In this paper, we deal with a selection of 10 most relevant publications. A more detailed evaluation of the papers is reserved for an extended version of this article.

Bellamy et al. [4] present a Python toolkit for algorithmic fairness, AI Fairness 360 (AIF360) at https://github.com/ibm/aif360, an extensible architecture, released under an Apache v2.0 license. It includes a comprehensive set of fairness metrics for datasets and models, explanations for these metrics, and algorithms to mitigate bias in datasets and models. The tool provides an interface to seven popular datasets concerning Census, Income, Credit, Recidivism, Marketing, and three versions of Medical Expenditure Panel Surveys (MEPS), a set of large-scale surveys of families and individuals.

The paper starts with a section on terminology and introduce the following concepts: A protected attribute, such as origin/ethnicity, gender, professional group membership or religion, creates a partition of a population (in relation to place and time) whose subsets are equal in terms of application-specific performance in the sense of fairness. Group/individual fairness aims to treat groups/individuals with protected characteristics similarly with similar outcomes when applying an algorithm or decision procedure. A privileged value of a protected attribute refers to a group or individual that has a systematic advantage.

Bias prevents fairness. Lack of fairness can be equated with undesirable biases that systematically favor privileged groups/individuals and systematically disadvantage unprivileged groups/individuals. A fairness metric allows the quantification of undesirable biases in training data for groups/individuals or models. A bias reduction algorithm is a method for reducing undesirable biases in data sets or models.

Chen [5] provides a comprehensive overview of the fairness issues in AI systems through practical applications, e.g., social administration, and focuses on bias analysis and fairness training. The assessment of fairness across the individual stages of an AI-supported decision-making process is worth to be considered in this context, starting with careful planning. In the medical field, this begins with the selection of cohorts, the consideration of databases from different national or international health organizations regarding standards and similar studies and the handling of epistemic uncertainty, the lack of data, their unconscious or conscious exclusion, the quality of the calibration of the models and their stability or sensitivity. Here, individual fairness means that similar patients receive equivalent treatments, which requires similarity metrics.

A similarity definition is missing in [5] and could be based on various distance notions and measurements applied to points, sets, text, or images. Classification uses statistical evaluation metrics based on quality criteria such as accuracy and efficiency sta-

tistical analysis and metrics such as F-score, precision, recall, sensitivity and specificity, AUC (Area under the curve), ROC (Receiver operating characteristic), or, in the best case, a clear individual assignment to non-overlapping groups of patients.

The paper [5] further assesses concepts of fairness and discrimination in the digital transformation of public health, the application of AI and ML algorithms in the medical field and their impact on patient health from the perspective of the stakeholders, patients, and experts involved. The authors discuss ethical and legal considerations such as data protection, responsibility, accountability, transparency and explainability in AI as is also addressed in [11].

Poor models lead to classification errors, e.g., when assigning people to risk groups, which have a direct negative impact on those affected. Fairness must be evaluated alongside accuracy and requires predictive models to not unfairly disadvantage specific demographic groups. To evaluate fairness in prediction models for development of psychosis and functional outcome. The authors of [6, 18] evaluated relevant fairness aspects for the demographic attributes 'gender' and 'educational attainment' and compared them with the fairness of clinicians' judgements. In general, it is preferable to use models which, when validated, result in the correct classification for each individual in the cohort, or at least correct lower or upper limits when it comes to risk prediction.

Educational bias was present in algorithmic and clinicians' predictions, assuming more favorable outcomes for individuals with higher educational level (years of education). [22] presents normative recommendations in the form of a statement on Fairness of AI Recommendations in Healthcare (FAIR), which summarizes best practices. With a comprehensive exploration of bias, fairness, transparency, and accountability, it guides readers through the intricate web of ethical considerations.

However, it should be noted that the risk of disease may well depend on gender or origin, as is shown by the occurrence of harmful gene mutations in ethnic groups with founder effects [15] or more frequent occurrence of normal-tension glaucoma in females [29, reference 80].

Linardatos [13] lists, analyzes and compares interpretability methods for explaining deep learning models (such as black box models), compiles tables that link explainable AI (XAI) technologies with application companions, models and methods, fairness issues, and also names unsuitable approaches. "Systems whose decisions are not easily interpretable are difficult to trust, especially in areas such as healthcare or self-driving cars, where moral and fairness issues naturally arise." Several techniques for controlling discrimination and removing the bias from machine learning models are presented.

In the sequel, we will discuss aspects of algorithmic fairness, give some examples of online risk prediction models and briefly characterize them. A comparison of the results of risk calculators and related studies is possible only if similar patient groups are medically treated according to the same procedures and similar protocols, and if the measurements use the same procedures, model parameters, and concern overlapping periods. For example, measurements are undertaken for both eyes several times in certain times under defined conditions using the same procedures, and a statistical model adjustment and calibration have been carried out.

Epidemic uncertainty in the results occurs if measured values are missing due to the absence of test subjects or if personal data is missing or incorrectly collected. With the help of interval arithmetic and the interval-based Dempster Shafer theory (DST), these can be considered directly in the risk calculations.

2 A Critical Debate on General Algorithm Fairness

A score-based tool metric is part of a complex audit procedure and is made up of individual numerical values (scores) that are assigned according to a specific description of requirements. They allow the assessment of the risk model, the factors and classes, the patient groups and individual risks, the tool (software, middleware, and hardware) and the data management. An evaluation of the interaction between the participants also concerns patient care and communication with experts, the cooperation of the groups involved and is supplemented by process validation over appropriate periods of time, which checks the assignment of patients to (risk) groups over a longer period of time and corrects it, if necessary, but also compares it with the findings from other models, patient groups and organizations. In this article, we limit ourselves to online medical tools and their interfaces to databases, patients and experts. There are various reasons for this.

Requirements such as correct input of medical parameters in the specified arithmetic formats must be supported and checked by the implemented model for correctness and completeness in the background. Methods for determination change over time, are only convertible or comparable with each other to a limited extent or, in the case of repeatedly collected health parameters in a given period, must be transparently averaged and assigned to a specified value range, which ultimately means that reliable operation and use of the tools should only take place in four-eyes operation.

The paper by Loftus et al. [14] deals with current approaches to algorithmic fairness using causal reasoning. As the most important conclusion, the authors see a fundamental disagreement on the way to a generally valid definition of algorithmic fairness. Are algorithmic decision-making procedures fair if they always make similar decisions for similar individuals, or if they make advantageous decisions for all groups at the same rate? As a consistent rule, criteria from different value systems, cannot be fulfilled simultaneously. In addition, algorithms use data and are part of one or more of many models. Thus, disadvantages can also arise in data collection, selection and depends on the scope of use, and this applies equally to input and results. As a conclusion, the authors state that "machine learning algorithms can unwittingly perpetuate or create discriminatory decisions that are biased against certain individuals".

Green's paper [7] presents a literature review on algorithm fairness, which concludes with the following words: "Although no mathematical definition of algorithmic fairness fully encapsulates the philosophical notion of fairness or justice, each definition captures a normatively desirable principle".

Instead of treating fairness as a technical attribute of algorithms, the author introduces substantive algorithmic fairness that focuses on whether and how algorithms can promote equity in practice.

Whereas "formal algorithmic fairness" is a decision-making process that aligns with formal equality (which emphasizes equal treatment for individuals based on their attributes or behavior at a particular decision), substantive algorithmic fairness focuses on whether and how algorithms can promote equity in practice: "First, reduce the upstream (social) disparities that feed into decision-making processes. Second, reduce the down-stream harms that result for those judged unfavorably within the decision-making process."

Weinkauf [26] states that "algorithmic fairness shares in the basic properties of fairness, but it also differentiates itself from the ethical concept of fairness by focusing on the treatment of individuals and groups within the context of an AI/ML model"

He recommends measures that can be applied to help achieve algorithmic fairness:

- Ensure that training data is balanced and representative of the population the ground truth is objective
- Ensure that features equally predict the target variable across groups
- Set the threshold to a value that satisfies your fairness criteria
- Use a separate threshold for each group.

In conclusion, as is stated by the author, mathematical notions don't address the ethical implications of the actual task an AI/ML model performs. Algorithmic fairness consists of multiple definitions that are generally incompatible, its results can be manipulated, and it cannot evaluate its effects.

The establishment of a generally valid metric for algorithmic fairness is controversial in the literature. Depending on the origin of the experts, various reasons are given for this, all of which are similar in that logical constructs and ethical maxims cannot be reconciled, as they are based on incompatible axioms or normative systems.

Particular difficulties also arise from the fact that there are extremely diverse approaches to the algorithmic classification of patients and their assignment to similar and dissimilar groups, which combine statistical measures or evidence-theoretical approaches such as DST or IDST with interval calculation to include epistemic uncertainty. Undoubtedly, the best solution is the most complicated one, namely to consider an online risk prediction tool as fair if all patients have a disease progression that corresponds to the predicted risk group at the end of the period.

3 Biases in Medical Tools for Risk Prevention

Following [24], we will examine biases in medical tools for risk prevention that affect data selection cohorts, interaction with experts and their collaboration. From this perspective, we call a computer-based medical prediction tool fair if it meets the following requirements.

a) The selection, composition, and follow-up of the cohorts used for modeling and validation of the prediction were accurately described, and the data were classified, processed and made accessible according to international quality standards

b) The modeling and result generation, including the theoretical foundations, technologies, architectures used and their validation, are described in comparison with existing tools and validated according to international quality criteria and quality measures, considering aleatory and epistemic uncertainty. To this end, it is necessary to follow the test subject group at least over the period of prediction, but preferably over a much longer period, and to adapt the tools for other test subject groups and further medical treatment, which particularly concerns progresses in the clarification of genetic influences.

c) Fairness means that subgroups in the cohort that have special additional (unique) characteristics or whose numbers result in a higher uncertainty factor or require special treatment methods are not disfavored [20].

d) Preferably, validated realistic lower and upper limits are set for the risk classes, within those limits a minimum standard of care is guaranteed [3].

e) The treatment of clinical conditions is monitored by a panel of experts over the period established at the time of diagnosis and is adjusted accordingly as relevant new information becomes available. [8, 10].

The discussion of the concept of fairness depends greatly on the specialist community in which it is used–while economics is about assessing the creditworthiness of a person or institution fairly, selection interviews for a new position are about not disadvantaging applicants because they belong to a particular group. Computer science is concerned with algorithmic fairness, in decision-making problems or the comprehensible use of AI and its results.

Classification errors happen when assigning people to their risk groups, due to lack of data and poor models, missing calibration, or uncertainty issues. Quality criteria and metrics guarantee the correct classification for each individual in the cohort, or at least correct and tight lower and upper bounds for risk prediction, while game theory is concerned with fair rules, utility functions and equilibrium that are included in the decisions of individuals and the associated motivations for reward or punishment. The assessment of fairness across the individual stages of an AI-supported knowledge creating and decision-making process is also important from the perspective of the political institutions and described in German Standardization Roadmap Artificial Intelligence in November 2020 at https://www.dke.de.

4 Case Study in Fairness in the Use of Medical Tools

Fairness in digital healthcare is based on the World Medical Association. Declaration of Geneva. World Medical Association; 1983: "A member of the medical profession will not permit considerations of age, disease or disability, creed, ethnic origin, gender, nationality, political affiliation, race, sexual orientation, social standing or any other factor to intervene between my duty and my patient".

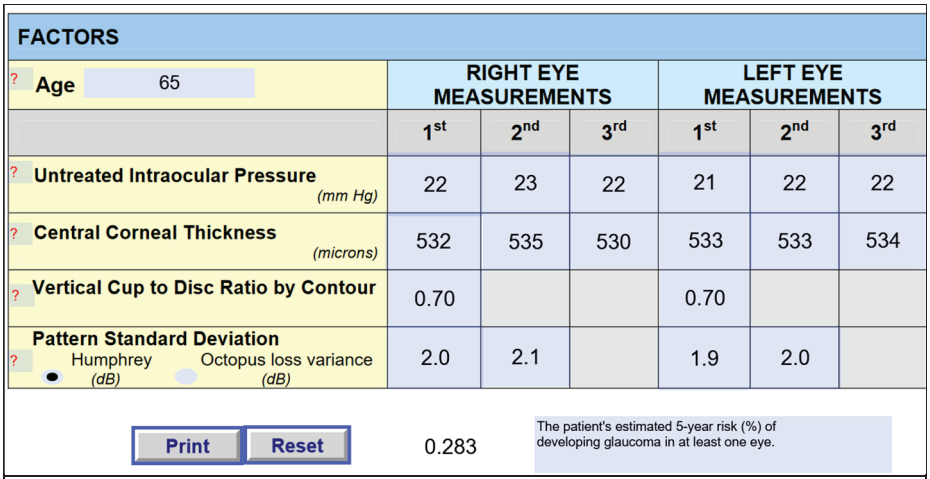
Biases of healthcare in digital models of diagnostics, prevention and treatment concern content, i.e., data, algorithms, and knowledge creation using AI and ML, patient cohorts, medical experts and their interaction and collaboration.

In a recent work [15] we presented online tools that calculate the individual and familial risk for the occurrence of pathogenic variants in BRCA1 (Breast Cancer gene) or BRCA2 genes with impact on early breast (BC) and ovarian cancer (OC) disease, and for estimating the 5-year risk that an individual with ocular hypertension will develop Primary Open Angle Glaucoma (POAG), the leading global cause of irreversible blindness, the angle between the iris and cornea remains open, and drainage does not work properly [3, 15, 28].

The forms collect information about the individuals, their specific disease patterns, medical examination results, and the lifestyle of the proband and his/her relatives. Data and model quality and cross-cutting issues such as uncertainty and usability, and fairness will be addressed in the context of work in which the authors have been involved. Finally, important requirements for online risk prediction tools are formulated, also considering aspects of fairness. An example of Glaucoma disease is resumed in Figure 1. The tool uses a log-log regression model that cannot be easily adapted to specific cohorts or ethnicities. In this respect, the reference to possible errors in the risk results is justified, but not conclusive for the groups concerned.

The modeling and result generation, the theoretical foundations, technologies, architectures used are described in comparison with existing tools and validated according to international quality criteria and metrics, considering aleatory and epistemic uncertainty. To this end, it is necessary to follow the test and control subject groups at least over the period of prediction, but preferably over a longer period, and to adapt the

results for other test subject groups and further developed investigation methods, which particularly concern progress in the clarification of the genetic impact [12, 15, 16, 19].



FACTORS						
Age 65	RIGHT EYE MEASUREMENTS			LEFT EYE MEASUREMENTS		
	1st	2nd	3rd	1st	2nd	3rd
Untreated Intraocular Pressure (mm Hg)	22	23	22	21	22	22
Central Corneal Thickness (microns)	532	535	530	533	533	534
Vertical Cup to Disc Ratio by Contour	0.70			0.70		
Pattern Standard Deviation: Humphrey (dB) / Octopus loss variance (dB)	2.0	2.1		1.9	2.0	

Fig. 1. The continuous method WPOAG for estimating the 5-year risk R of developing POAG $R = 1 - 0.911^{\exp(0.2792\cdot((age-56)/12 + (IOP-25)/3) + 0.75566\cdot(573-CCT)/42 + 0.6262\cdot((PSD-1.9) + 10(VCD-3.9)/3.5))}$ [17]

Risk Calculator	Scientific study basis: Ocular Hypertension Treatment Study (OHTS) [9] and the European Glaucoma Prevention Study (EGPS)
Content dedication and users	5-year risk that an individual with ocular hypertension will develop POAG; clinicians and patients
Scientific basis	DOI: 10.1016/j.ophtha.2006.08.031–Validated prediction model for the development of POAG in individuals with ocular hypertension
Information needed for use	1) Age 2) Vertical cup/disc ratio (VCD) by contour 3) Intraocular pressure (IOP, mmHg) 4) Central Corneal Thickness (CCT, mm) 5) PSD Visual field pattern standard deviation (PSD, dB) and other parameters as (corrected Octopus) loss variance (OLV, dB)–IOP (3 meas.) per eye using Goldmann applanation tonometry), CCT using an ultrasound pachymeter (3 meas. per eye), PSD using any of the following (2 meas. per eye, Humphrey full threshold, SITA standard 30-2 or 24-2, LV from Octopus 32-2
Methods	For the Continuous Method enter actual data for the patient's eyes, for the Point System select the range for the patient's age and average of the multiple meas.
Results	Both methods give similar, but not identical results
Limitations and cautions	There is no guarantee that the predicted risk is accurate for individual patients. It is not clear whether these models predict progression of established disease.
Parameter ranges	30y≤a≤80y; IOP range 20–32 mm Hg; CCT 475–658 microns; VCD 0–0.8; PSD 0.5–3.0 dB
About	Information for participants–ocular hypertension and glaucoma–POAG
Copyright: WU 2006	The predictions derived using these methods are designed to aid, but not to replace, clinical judgment.
Department of Ophthalmology & Visual Sciences, Washington University School of Medicine	

Fairness also means that subgroups in the cohort that have special additional/unique characteristics, whose numbers result in a higher uncertainty factor or require special treatment methods, are not disfavored.

Validated lower and upper bounds are specified for the risk classes, a standard of care guaranteed.

The patient's treatment and the use of suitable technologies are continuously monitored by a panel of experts from the time of initial diagnosis and adjusted accordingly if new insights become available.

In general, it can be said that these requirements are met in the BRCA and glaucoma risk tools under consideration, but with the following restriction. The online form presented to the patient is only used to collect their data and assign them to a risk group. It is usually provided with instructions on how to use it, how to interpret the results, information on advice centers and their services and on version development, the disclaimer and accompanying scientific literature [1, 17].

In addition, the tools are validated in scientific (meta-)studies and compared or critically examined in terms of modeling, algorithms and accuracy as well as applicability for different population groups and treatment periods. In this respect, they are useful for the layperson if medical or nursing expertise can be directly involved [2, 8, 10, 12, 17, 21]. Kuchenbaecker et al. [10] examine some selection and survival biases and their reasons.

Kass et al. [8] aimed to determine the cumulative incidence and severity of POAG after 20 years of follow-up among participants in the OHT study.

To determine whether bias occurred in re-participation, the authors compared participants who re-enrolled by randomization group and by baseline clinical and demographic data with those who did not re-enroll.

To guarantee fairness in AI-based risk assessment online tools, we would like to derive some minimum requirements [17] as a score-based bias metric: with a range of 0–15pts.

- Model or classification algorithm inaccurate → exit
- General information: Adequate information about their purpose, their operation, and the handling of the output results is needed (1 pt.)
- Risk factors: Comprehensible and fair information must be available for each particular question: What kind of information is expected about the individual's demographics, lifestyle, health status, prior examination results, and family disease history? (2 pts.)
- Risk classes: Depending on the disease pattern, examination outcomes, and patients' own medical lab samples (e.g., biomarkers), patients are assigned to a risk class that is clearly described. If terms such as high risk, low risk etc. are used, transition classes should be provided to avoid assigning similar individuals to dissimilar classes (2 pts.)
- Risk model includes risk factors for special groups (1 pts.)
- Assistance: Questionnaires could be completed in a collaborative manner (1 pt.)
- Data handling: Data and results must also be at disposal over a longer period of time, cross-cutting requirements such as data protection, privacy, and security are respected (2 pts.)
- Expert assistance: Involvement of treating doctor; experts should be given references to relevant literature on data, models, and algorithms, validation and follow-up (2 pts.)
- Result consequences: Outcomes should be designed in such a way that users are provided with appropriate counseling and help service depending on their allocation to a risk class and reference to effects of various sources of uncertainty (3 pts.)

- Arbitration boards and mediation procedures in the case of disputes are provided (1 pt.)

Examples: Score-based Laroche Glaucoma Calculator

Patient 50/51y, 1/2 pts. Mean IOP 18/19 2/3 pts. CCT 499/500 1/2 pts. → 4/7 pts. Low/high risk. Inaccurate model, since small changes in data result in a change in the risk class: the tool should provide an intermediate class for these cases.

WPOAG point system risk calculator: patient 64/65y, Mean IOP 24 mmHg, CCT 550, Mean VCR 0.4, PSD 2.4 results in 12/13 pts. (intermediate /high risk).

In a recent paper [15] "Validation of Risk Assessment Models for Breast and Ovarian Cancer-Related Gene Variants", algorithmic fairness was achieved for two risk metrics that provide mutual statements about their suitability and about quality criteria such as performance, accuracy and consistency for the risk models, genetic counseling tools and comparative surveys. The universal metrics reproduce the mean probability of a BRCA1/2 mutation determined in five large studies with cohorts from different ethnicities and risk classes, suggesting their suitability for further studies as well. Firstly, the study results were correctly calibrated, and secondly, the proposed metrics are universally applicable.

The models presented in [15] for calculating risk probabilities are based on the DST, which introduces basic probability assignments with masses m for the individual disease variants and uses their combined occurrences to calculate an overall risk. DST provides an explicit method that is very flexible and always comprehensible, thanks to further basic probability assignments (BPAs) with additional weights and the inclusion of uncertainty via interval arithmetic. In addition, BPAs for different patient groups, such as patients and their relatives, can also be combined using Dempster's rule. Further fields of application for DST are regression and prediction (cf. [15, p. 112]).

5 Conclusion

Changes and tightening in the definition of risk classes over the past 20 years, advances in imaging techniques and reporting systems, gene panel testing, and use of related genomic and biomarkers have a major impact on the models and algorithms underlying Cancer or Glaucoma Risk Calculators. Mixed gender and ethnicities in cohorts, lack of data on patients, their disease and family histories, different standards in digital examinations and polygenic tests performed may lead to highly variable results and non-comparability of risk calculators because of epistemic uncertainty, while the authors cited in the references focus mainly on aleatory uncertainty. There is still no consistent consideration of fairness issues like general algorithmic metrics for explainable AI-based software [19, 22, 23, 25, 29]. Further work concerns standardizing the development of evaluation frameworks for assessing fairness and biases in medical risk tools, to introduce comprehensible criteria catalogs for reliable fairness analytics and to apply them to known risk prediction online software in analogy to the approach in our paper [27] dealing with visual analytics.

Acknowledgement. The research was initially presented by the first author at a workshop on the dangers of discrimination through artificial intelligence held at the University of Cologne from March 13 to 15, 2024 at the invitation of Professors Dr. Wilfried Hinsch and Dr. Sven Nyholm. We thank our anonymous reviewers for their helpful and competent hints.

References

1. American Academy of Ophthalmology: Risk Calculators for Primary Open-Angle Glaucoma. Oct. 10, 2019. https://www.aao.org/education/interactive-tool/risk-calculators-primaryopen-angle-glaucoma
2. Amir, E. et al.: Assessing Women at High Risk of Breast Cancer: A Review of Risk Assessment Models. JNCI **102**(10), 680–691 (2010)
3. Auer, E., Luther, W.: Uncertainty Handling in Genetic Risk Assessment and Counseling. Journal of Universal Computer Science **27**(12), 1347–1370 (2021)
4. Bellamy, R. et al.: AI Fairness 360: An Extensible Toolkit for Detecting and Mitigating Algorithmic Bias. IBM Journal of Research and Development **63**(4/5), 4:1–4:15 (2019) DOI:10.1147/JRD.2019.2942287
5. Chen, P., Wu. L., Wang, L.: AI Fairness in Data Management and Analytics: A Review on Challenges, Methodologies and Applications. Applied Sciences **13**(18):10258 (2023)
6. El Azab, S., Nong, P.: Clinical algorithms, racism, and "fairness" in healthcare: A case of bounded justice. Big Data & Society **10**(2), 13p. (2023)
7. Green, B.: Escaping the Impossibility of Fairness: From Formal to Substantive Algorithmic Fairness. Philosophy & Technology 35:90, 32p. (2022)
8. Kass, M. A. et al.: Assessment of Cumulative Incidence and Severity of Primary Open-Angle Glaucoma Among Participants in the Ocular Hypertension Treatment Study After 20 Years of Follow-up. JAMA Ophthalmol. **139**(5), 1–9 (2021) https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8050785/?report=printable
9. Karmel, M. Glaucoma: Calculating the Risk. EyeNet Magazine (2024) https://www.aao.org/eyenet/article/glaucoma-calculating-risk
10. Kuchenbaecker, K. B., Hopper, J. L., Barnes, D. R. et al.: Risks of breast, ovarian, and contralateral breast cancer for BRCA1 and BRCA2 Mutation Carriers. JAMA **317**(23), 2402–2416 (2017) DOI: 10.1001/jama.2017.7112.
11. Kumar, R., Joshi, A., Sharan, H.O., Peng, S.-L., Dudhagara, C.R.: The Ethical Frontier of AI and Data Analysis, IGI Global 2024
12. Laroche, D., Rickford, K., Mike, E. V., Hunter, L., Ede, E., Ng, C., and Douglas, J.: A Novel, Low-Cost Glaucoma Calculator to Identify Glaucoma Patients and Stratify Management. Hindawi J. of Ophthalmology, Article ID 5288726, 6 p. (2022)
13. Linardatos P, Papastefanopoulos V, Kotsiantis S. Explainable AI: A Review of Machine Learning Interpretability Methods. Entropy **23**(1) 18 (2021)
14. Loftus, J. R., Russell, C., Kusner, M. J., and Silva, R.: Causal Reasoning for Algorithmic Fairness. arXiv e-prints1805.05859, 5 (2018)
15. Luther, W.: Validation of Risk Assessment Models for Breast and Ovarian Cancer–Related Gene Variants, in Hajian, A., Baloian, N., Inoue, T., Luther, W. (eds.): Data Science, Human-Centered Computing, and Intelligent Technologies. Logos, Berlin, 106–111 (2022)
16. Owens, D. K.: Risk Assessment, Genetic Counseling, and Genetic Testing for BRCA-Related Cancer: US Preventive Services Task Force Recommendation Statement. JAMA **322** (7), 652–665 (2019) DOI: 10.1001/jama.2019.10987
17. Risk Calculators for POAG (OHTS/EGPS). https://ohts.wustl.edu/risk/
18. Şahin, D. et al.: Algorithmic fairness in precision psychiatry: analysis of prediction models in individuals at clinical high risk for psychosis. British Journal of Psychiatry **224**(2), 55-65 (2024) DOI 10.1192/bjp.2023.141
19. Shaw, J.: OHTS: 20 Years of Follow-up Data on POAG. EyeNet Magazine June 2021 https://www.aao.org/eyenet/article/ohts-20-years-of-follow-up-data-on-poa
20. Siegfried, C. J., Shui, Y. B.: Racial Disparities in Glaucoma: From Epidemiology to Pathophysiology. Mo Med. **11**(1), 49–54 (2022)
21. Tham, Y. C., Li, X., Wong, T. Y., Quigley, H. A., Aung, T., Cheng, C. Y.: Global Prevalence of Glaucoma and Projections of Glaucoma Burden Through 2040: A Systematic Review and Meta-Analysis. Ophthalmology **121**(11), 2081–90 (2014)

22. Tirendi, S., Domenicotti, C., Bassi, A. M., Vernazza, St.: Genetics and Glaucoma: The State of the Art. Frontiers in Medicine **10**(12):1289952 (2023)
23. Topouzis, F. and Giannoulis, D.: Understanding the Genetics of Glaucoma. Ophthalmology Times Europe **17**(4) May (2021)
24. Ueda, D., et al.: Fairness of artificial intelligence in healthcare: review and recommendations. Japanese Journal of Radiology **42**(1), 3–15 (2024)
25. Wang, Z., Wiggs, J. L., Aung, T., Khawaja. A. P., Khor, C. C.: The Genetic Basis for Adult-Onset Glaucoma: Recent Advances and Future Directions. Progress in Retinal and Eye Research 90, Paper ID 101066 (2022)
26. Weinkauf, D.: Privacy Tech-Know blog: When worlds collide –The possibilities and limits of algorithmic fairness. https://www.priv.gc.ca/en/blog/20230405_02/
27. Weyers, B., Auer, E., Luther, W. The role of Verification and Validation Techniques within Visual Analytics. JUCS **25**(8), 967–987 (2019)
28. World Health Organization: Blindness and Vision Impairment. Key Facts, 10 August 2023 https://www.who.int/news-room/fact-sheets/detail/blindness-and-visual-impairment
29. Zukerman., R., Harris, A., Vercellin, A. V., Siesky, B., Pasquale, L. R., Ciulla, T. A.: Molecular Genetics of Glaucoma: Subtype and Ethnicity Considerations. Genes (Basel) **12**(1) 55 (2021)

Exploring Design Aspects of an AI-supported Farming Platform

Arsen G. Mikayelyan[0009-0009-6936-2396] and Ashot N. Harutyunyan[0000-0003-2707-1039]

ML Lab, Yerevan State University, 0025 Yerevan, Armenia
arsen.mikayelyan@ysu.am, harutyunyan.ashot@ysu.am

Abstract. We are exploring design aspects and elements of a comprehensive platform for smart agriculture. Focusing on the main concepts and an earlier prototype functionality, we propose the related vision on building such a solution to modernize farming practices through the integration of cutting-edge artificial intelligence (AI) approaches. The platform will be consisting of a suite of innovative features aimed at optimizing crop cultivation, irrigation management, and strategic planning for agricultural enterprises and regulatory actors.

Keywords: Intelligent agriculture, ML analytics of agronomic data, Gen AI and LLMs, real-time recommendations for farming.

1 Concepts and System Design

As in all spheres of human activities and domains of business, modern agriculture as well increasingly relies on data-driven decision making, real-time analytics, and automated management. Prior works on using AI in the agriculture include various studies and technology solutions. However, the review paper by Spanaki et al. [1] indicates that the disruptive potential of AI in the agricultural sector in terms of research and operations are still in infancy. A recent paper [2] focuses on opportunities and challenges that AI-driven approaches imply for a sustainable development in African continent. Vendors like IntelinAir [3] apply image processing techniques to effectively monitor health of agronomic fields. There are also various data sets [4]-[8] that can be helpful for building ML models in this domain.

In our study, we think more of an approach which is “platform”-ic in nature, comprising key features like an AI-driven assistant for planting guidance, an intelligent irrigation scheduling system with real-time monitoring capabilities, and personalized diagnostics for crop cultivation. Through the utilization of state-of-the-art ML algorithms, as well as Gen AI capabilities and Large Language Models (LLMs), the platform might provide dynamic conversational interfaces and a question-answering system, data analysis, and advanced statistical modeling to empower users with actionable insights.

In a full implementation scenario, the hierarchical backend service (Fig. 1) ensures seamless scalability, accommodating the diverse needs of multi-group users while maintaining robust performance and reliability. The platform's user-friendly interface enables farmers to visualize water lines, plot dimensions, and crop-specific cultivation

information, facilitating informed decision-making and maximizing agricultural productivity. We enable the system's features accessible through mobile devices

By democratizing access to expert-level insights and automation tools, this platform represents a significant advancement in precision farming technology. Its ability to optimize irrigation schedules, provide personalized recommendations, and streamline agricultural operations has the potential to enhance sustainability, increase crop yields, and improve food security on local and larger scales.

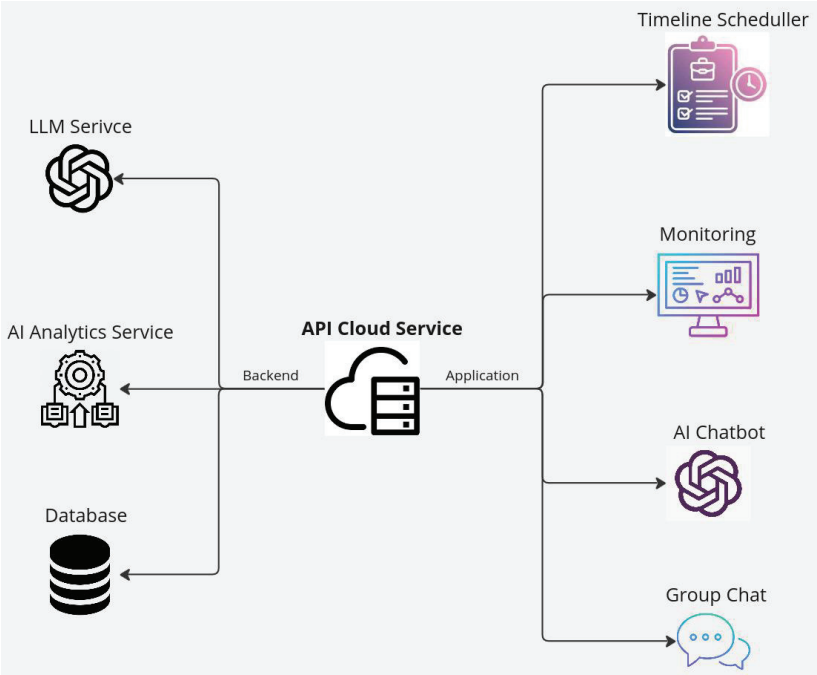


Fig. 1. Proposed architecture of the system.

The specific capabilities of the proposed system include:

- **User Registration and Personalized Dashboard.** Upon registration, users gain access to a personalized dashboard tailored to their agricultural needs. The dashboard provides an interactive map view of the user's agricultural area, showcasing plot dimensions and irrigation infrastructure. Detailed histories of planted crops and watering schedules will be easily accessible, enabling users to track progress and make informed decisions.

- **Watering Plan Scheduling and Coordination**. Users can effortlessly schedule watering plans using the platform's interface, ensuring optimal irrigation management. A centralized reservation system allows users to view and coordinate watering schedules for irrigation, promoting resource optimization and collaboration.

- **Data Analysis and AI Guidance.** ML algorithms, classification and regression models analyze user-specific data, including crop types and soil characteristics, to provide personalized guidance. These algorithms offer tailored recommendations for planting processes, resource allocation, and strategic decision-making based on data insights.

- **Real-Time Monitoring and Cost Analysis:** Modern monitoring tools (such as those for collecting time series data) provide real-time insights into watering plant frequency and associated costs, empowering users to optimize resource usage and budget effectively.

- **LLM Assistant and Retrieval Augmented Generation (RAG).** A Gen AI-based assistant (e.g., a question-answering system) harnesses user historical data and employs RAG techniques to deliver personalized guidance and support using up to date and relevant data. The assistant proactively alerts users to potential issues, offers Chabot functionality with LLMs, and facilitates seamless communication for immediate problem solving.

- **Community Communication Platform:** The platform serves as a vibrant community hub where users can engage in discussions, share insights, and exchange best practices. Through integrated chat features and discussion forums, users foster collaboration, seek advice, and celebrate successes within the agricultural community.

2 Training Models for Recommendations

An exemplary situation that a farmer faces is when she needs to decide about the most effective crop type for a given land. Using the training data set in [7] with seven features on the soil characteristics for 4.4K sample records and the corresponding label (sort of plant), the platform produces recommendations based on predicted outcomes for a given soil parameters. In this study, four ML models were employed: Decision Tree, Support Vector Machine (SVM), Random Forest, and K-Nearest Neighbors (KNN). Each model was trained on a dataset comprising soil characteristics including data specific to the Armavir region [8]-[9], along with corresponding crop types. The performance of each model was assessed using metrics such as accuracy and macro-average F1 score (Figs. 2, 3).

	Model name	Best score
0	decision tree	0.986667
1	svm	0.986061
2	random_forest	0.993939
3	k classifier	0.976364

Fig. 2. Test results of ML models trained.

The Random Forest model demonstrated the highest accuracy among all the models, achieving a remarkable score of 99.39% on the validation set with 10 folds. This indicates its superior ability to classify crop types based on the given soil characteristics including also Armavir region. Additionally, we didn't observe divergence in models performances before and after inclusion of the data from Armavir region which comprises approximately 10% of the overall data set.

	precision	recall	f1-score	support
potato	0.83	0.95	0.89	21
cucumber	0.83	1.00	0.91	15
tomato	0.96	0.96	0.96	74
corn	0.94	0.96	0.95	53
carrot	0.93	0.93	0.93	70
eggplant	0.98	0.87	0.92	53
peas	0.95	0.90	0.92	40
onion	0.89	0.93	0.91	121
apple	0.94	0.91	0.93	35
cherry	0.88	0.87	0.88	155
strawberries	0.87	0.93	0.90	58
wattermelon	0.94	0.96	0.95	69
grape	0.93	0.92	0.92	454
accuracy			0.92	1218
macro avg	0.91	0.93	0.92	1218
weighted avg	0.92	0.92	0.92	1218

Fig. 3. Test results on Armavir data of ML models across types of crops.

The high accuracy and F1 score across all models suggest that the AI-guided recommendations derived from these models can be reliable for making informed decisions related to crop selection, resource allocation, and strategic planning in agriculture.

3 Conclusion

With this short paper we specify the larger vision behind the proposed system which needs to be realized based on an initial prototype. That will be demoed at the workshop. We believe that such a platform solution represents a paradigm shift in farming practices, harnessing advanced AI techniques to address modern agricultural challenges. Through its features including Gen AI-based guidance, the platform empowers users to optimize resource allocation and maximize yields. Its collaborative features might foster knowledge sharing, innovation, higher productivity and sustainability within the agricultural community.

References

1. Spanaki, K., Sivarajah, U., Fakhimi, M., Despoundi, S., Irani, Z: Disruptive technologies in agricultural operations: a systematic review of AI-driven AgriTech research. Annals of Operations Research **308**, 491–524 (2022): https://doi.org/10.1007/s10479-020-03922-z
2. Gikunda, K.: Harnessing artificial intelligence for sustainable agricultural development in Africa: Opportunities, challenges, and impact: https://arxiv.org/html/2401.06171v1
3. Actionable Intelligence from Aerial Data: https://www.intelinair.com/
4. Agricultural Data for Rajasthan, India (2018-2019) https://www.kaggle.com/datasets/suraj520/agricultural-data-for-rajasthan-india-2018-2019
5. Crop Recommender Dataset with Soil Nutrients: https://www.kaggle.com/datasets/manikantasanjayv/crop-recommender-dataset-with-soil-nutrients/data
6. Crop Yield Prediction Dataset: https://datasets.omdena.com/dataset/crop-yield-prediction
7. Open Source Geospatial Content Management System: https://geonode.org/
8. The Armenian Soil Information System (ArmSIS): http://armsis.cas.am/layers/?limit=5&offset=0
9. Harutyunyan, S.S., Matevosyan, L.G., Ghukasyan, A.G., and Galstyan, M.H.: The system of soil protection and general balance of main nutrient elements in perennial plantations of semi-desert natural soil zone of Armenia: https://doi.org/10.15159/AR.22.039

An Explainable Clustering Algorithm using Dempster-Shafer Theory

As noted by *R. Valdivia, N. Baloian, M. Chahverdian, A. Adamyan, and A. Harutyunyan*, traditional clustering algorithms like K-Means or agglomerative clustering often lack interpretability, making it difficult for users to understand the underlying patterns and rules. To address this issue, the authors recommend a clustering algorithm that generates labels for data and uses the Dempster-Shafer classifier to create clear rules, ensuring interpretability for users.

Embedded Interpretable Regression using Dempster-Shafer Theory

N. Baloian, E. Davtyan, K. Petrosyan, A. Poghosyan, A. Harutyunyan, and S. Peñafiel suggest a method, termed Embedded Interpretable Regression, which segments continuous output variables into discrete classes and applies two stage procedures to obtain regression output. In the first stage, a Dempster-Shafer Theory-based classifier is trained to assign new instances to these predefined buckets. In the second stage a specific regressor is trained for each bucket to obtain predictions.

Improving the DSGD Classifier with an Initialization Technique for Mass Assignment Functions

Creating a better 'starting point' than the original work, *A. Tarkhanyan and A. Harutyunyan* realize a speedup in training for the Dempster-Shafer Interpretable Classifier, which makes it converge faster without a significant loss in confidence and accuracy. Quality criteria concern the rule's uncertainty, confidence level, and representativeness

An Empirical Analysis of Feature Engineering for Dempster-Shafer Classifier as a Rule Validator

Explainable AI (XAI) has emerged as a crucial field, aiming to shed light on these black-box models and enhance trust in their outputs. *A. Baloyan, A. Aramyan, N. Baloian, A. Poghosyan, A. Harutyunyan, and S. Peñafiel* introduce an explainable classifier using Dempster-Shafer (DS) theory to validate AI-generated rule sets. DS theory combines evidence from various sources, handling conflicting information to identify reliable rules. The work shows that the DS-based classifier excels in learning numeric feature interactions (counts, differences, etc.) and performs well with class imbalance, but it struggles with imbalanced categorical data.

Interpretability of Machine Learning Models in the Insurance Sector

A. Sargsyan examines the challenges and methods associated with improving the interpretability of ML models, with a particular focus on their application in the insurance sector. The various stakeholders should understand how the model works and how their needs can be met with different approaches.

An Explainable Clustering Algorithm using Dempster-Shafer Theory

Ricardo Valdivia[1], Nelson Baloian[1][0000-0003-1608-6454], Maral Chahverdian[2], Aram Adamyan[2], Ashot N. Harutyunyan[3,4][0000-0003-2707-1039]

[1] Department of Computer Science, University of Chile, Santiago 8330111, Chile
[2] American University of Armenia
[3] ML Lab, Yerevan State University, 0025 Yerevan, Armenia
[4] Institute for Informatics and Automation Problems of NAS RA, 0014 Yerevan, Armenia

Abstract. Clustering is an unsupervised learning method aimed at identifying data sets with similar characteristics. The quality of a clustering model is often assessed by its validity rather than its accuracy, using measures such as the Rand Index and the Correlation Coefficient. Recently, there has been an increasing interest in creating not only valid but also interpretable clustering models. The proposed solution in this study involves a clustering algorithm that generates labels for data and using the Dempster-Shafer classifier, creates clear rules ensuring interpretability for users.

Keywords: Explainable ML/AI, Clustering, Classification, Dempster-Shafer Theory.

1 Introduction and Motivation

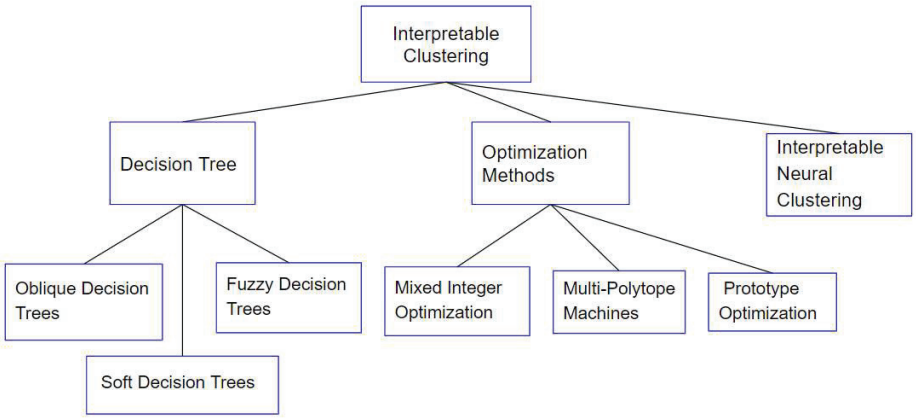
Interpretability [1,2] refers to the model's ability to enable a human user to understand the how and why behind the model's specific outcomes. Current clustering algorithms, like K-means, are favored for their simplicity and scalability, yet they are often viewed as "black box" due to their opaque results. This has led to a growing focus on understanding and interpreting clustering models, and in developing model explanation techniques, such as SHAP (SHapley Additive exPlanations) [8], to provide a clear understanding of how clustering results are produced. The proposed solution in this study involves the development of a clustering algorithm that generates labels for data and, using the DS (Dempster-Shafer) classifier [3], creates clear rules ensuring interpretability for users. The development occurs in two stages: selecting optimal labels for training and consolidating the clustering algorithm, including training and predicting with the DS classifier for each data point. The implemented DSClustering algorithm achieves an effective combination of clustering techniques with enhanced interpretation through the automatic generation of categorical rules and precise adjustments in the training process of the classifier. The algorithm stands out for its ability to provide reliable and comprehensible clustering results, enhancing transparency and trust in data-driven decision-making.

2 Literature Review

2.1 Prior Knowledge

There have been significant efforts in the development of interpretable clustering algorithms, employing various methods such as post hoc analysis of clusters and intrinsic interpretability approaches. Below you can find the most relevant and state of the art methods used for interpretable clustering. While many existing methods have used classification techniques such as decision trees our approach focuses on using the Dempster-Shafer Classifier algorithm which will be elaborated further in the paper.

Fig 1. The graph shows the categorization of Interpretable clustering. [9]



2.2 Prior Work

In the work by Peñafiel et al. [3], a classifier called Dempster-Shafer (DS classifier) has been designed using Dempster-Shafer's theory of plausibility [4]. This model has proven to be equally effective as other black-box models in terms of accuracy while also having the advantage of being easily interpretable. Interpretability in this context refers to the model's ability to provide simple and understandable rules used for predicting outcomes through machine learning. Due to the growing demand for clustering solutions that are both highly valid and interpretable, there is a need to continue developing methods that combine both features without seriously compromising the quality of any of them. Therefore, an interesting hypothesis from a research point of view is that it is feasible to generate a clustering algorithm that allows obtaining comparable results in terms of validity and interpretative ability compared to popular clustering algorithms such as k-means, DBSCAN, and agglomerative clustering but it is highly

interpretable at the same time. This approach could help users better understand the clustering process and in still greater confidence in the results obtained. The gap that this research aims to fill lies in the development of a clustering algorithm that provides a unique balance between validity and interpretability, a rare combination in current clustering methods. Unlike existing interpretable clustering algorithms, which often compromise validity, the proposed algorithm maintains high validity while providing simple rules for user understanding. It is also the only one that provides a measure of uncertainty for each cluster assignment. This dual approach increases user confidence and understanding of the underlying data patterns by addressing the need for reliable clustering solutions that are not only valid, but also transparent and understandable.

3 DSCLustering Algorithm

The main design principles of the proposed algorithm with explainability are as follows. First, a standard clustering process is performed on the data, so that each of data points is assigned to a cluster. Then, depending on the user's preference, the algorithm with the best Silhouette score or the most frequent (most repeated) cluster for each data instance is chosen. Then, considering each cluster as a class, an interpretable classifier (DS classifier developed in [3]) is trained with this labelled data, which generates rules to assign the samples to a certain class, so that we have the rules of how the discovered clusters are formed. Thus, it offers a novel approach to clustering that combines the best of two worlds: the reliability of the best-known clustering methods and the clarity of interpretable models provided by the DS classifier [3].

4 Experimental Results

To evaluate the results of the newly designed clustering algorithm, experiments were conducted assessing three important metrics relevant to clustering algorithms: the Silhouette score, Pearson correlation coefficient, and Rand index. The silhouette score is a measure of how similar an object is to its own cluster compared to other clusters and is calculated by taking into account the mean intra-cluster distance a and the mean nearest-cluster distance b for each data point. The silhouette coefficient for a sample is $(b - a)/\max(a, b)$. The values ranges from -1 to +1 where near + 1 means the data point is in the correct cluster and value near -1 means, the data point is in (a) wrong cluster.[5] The Rand Index computes a similarity measure between two clustering by considering all pairs of samples and counting pairs that are assigned in the same or different clusters in the predicted and true clustering.[6] RI = (number of agreeing pairs) / (number of pairs) which varies between 0 and 1. [7] Higher RI means that the clustering result is more similar to the ground truth clustering. These results were compared with three popular clustering algorithms and across three potential usage scenarios of DSClustering for users. This aims to answer the following questions:

- How does the clustering algorithm perform in terms of the validity of its results?
- Do the results of this algorithm compare favorably with those of popular classes for certain data types?

The Quality indices help us to evaluate clustering performance, however validation in specific industries such as medical examination and healthcare remains challenging. To overcome this issue, we should conduct extended analysis using domain-specific metrics that can further ensure that our clusters are clinically meaningful.

One of the main hypotheses that emerge is that the DS clustering algorithm should yield results similar to those of popular clustering algorithms. This is because it uses these algorithms as a foundation, particularly their outcomes, as expert experience to feed the necessary rules for predicting the labels. The results presented in Table 1 (DSC stands for DSClustering) confirms such an expectation. Moreover, the DSC methods show remarkable performance of quality scores on synthetic datasets of the distributions. It loses its effectiveness and consistency with real datasets which was expected in return of interpretability. Keep in mind that the using this algorithm allows the expert to feed rules which could further improve the accuracy and validate them.

Table 1. Comparison of metrics for different clustering methods on various datasets.

Dataset	method	Rand Index	Silhouette	Pearson
Uniform line	KMeans	0.968192	0.6092	-0.984111
	DBSCAN	0.0	0.0	0.0
	Agglomerative	0.626526	0.6008	-0.807947
	DSC Best Labels	0.444714	0.574601	0.478302
	DSC Most Voted label	0.252542	0.323917	0.679125
	DSC with number of cluster	0.984032	0.6090	0.992022
Uniform Rectangle	KMeans	1.0	0.4396	1.0
	DBSCAN	0.0	0.0	0.0
	Agglomerative	1.0	0.4396	1.0
	DSC Best Labels	0.503001	0.495606	-0.065668
	DSC Most Voted label	0.503082	0.495662	-0.064969
	DSC with number of cluster	1.0	0.43963	-1.0
Gaussian Distribution	KMeans	0.984032	0.6435	-0.992000
	DBSCAN	0.0	0.0	0.0
	Agglomerative	0.984032	0.64351	-0.992000
	DSC Best Labels	0.968192	0.4653	0.984031
	DSC Most Voted label	0.948585	0.475981	0.961466
	DSC with number of cluster	0.984032	0.64352	0.992000
Gaussian mix Distribution	KMeans	1.0	0.684170	-0.5
	DBSCAN	0.570609	0.6048	9.11e-18
	Agglomerative	1.0	0.684170	0.5
	DSC Best Labels	0.869085	0.558766	0.219600
	DSC Most Voted label	0.869085	0.558766	0.219600
	DSC with number of cluster	0.570609	0.6048	9.11e-18
Iris	KMeans	0.730238	0.552819	0.224350
	DBSCAN	0.520619	0.486034	0.367712
	Agglomerative	0.731199	0.554324	0.205441
	DSC Best Labels	0.546422	0.509248	0.750098
	DSC Most Voted label	0.546422	0.509248	0.750098
	DSC with number of cluster	0.695636	0.478250	0.126230
Wine	KMeans	0.371114	0.571138	-0.029116
	DBSCAN	0.0	0.0	0.0
	Agglomerative	0.368402	0.564480	0.572525
	DSC Best Labels	0.313358	0.370674	-0.226526
	DSC Most Voted label	0.276761	0.324729	-0.252627
	DSC with number of cluster	0.304176	0.584262	-0.673827
Breast Cancer	KMeans	0.519788	0.6972	-0.697773
	DBSCAN	0.0	0.0	0.0
	Agglomerative	0.390288	0.6899	0.742352
	DSC Best Labels	0.0	0.0	0.0
	DSC Most Voted label	0.0	0.0	0.0
	DSC with number of cluster	0.0	0.0	0.0

5 Conclusions

During the study, based on our experiments on various data sets we arrived at the following conclusions:

- The algorithm's ability to achieve high Silhouette scores across datasets indicates its effectiveness in identifying well-separated clusters, a critical aspect in many practical applications.
- The adaptability of DSClustering to different datasets, along with its versatility in handling both simple and complex data structures, was a notable strength. This adaptability was evident in its ability to generate meaningful and interpretable rules, even with complex datasets.
- A key advantage of the DSClustering algorithm was its interpretability. The use of rules provided clear guidelines for cluster assignment, enhancing transparency and user trust in the model

6 Acknowledgement

The research was supported by ADVANCE Research Grants from the Foundation for Armenian Science and Technology.

References

1. 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., & Herrera, F.: Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI, Information Fusion **58**, 82-115 (2020)
2. Ribeira, M.T., Singh, S., Guestrin, C.: Why should I trust you?: Explaining the predictions of any classifier. arXiv: 1602.04938v1 (2016)
3. Peñafiel, S., Baloian, N., Sanson, H., and Pino, J.A.: Applying Dempster–Shafer theory for developing a flexible, accurate and interpretable classifier, Expert Systems with Applications, 148:113262 (2020)
4. Shafer, G.: A Mathematical Theory of Evidence. Princeton University Press (1976)
5. Warrens, M.J., van der Hoef, H.: Understanding the Rand Index. In: Imaizumi, T., Okada, A., Miyamoto, S., Sakaori, F., Yamamoto, Y., Vichi, M. (eds): Advanced Studies in Classification and Data Science. Studies in Classification, Data Analysis, and Knowledge Organization. Springer, Singapore (2020) https://doi.org/10.1007/978-981-15-3311-2_24
6. Hubert, L., Arabie, P.: Comparing partitions. Journal of Classification **2**, 193–218 (1985)
7. Wikipedia, Rand Index, retrieved June 30, 2024, https://en.wikipedia.org/wiki/Rand_index
8. Rodríguez-Pérez, R., & Bajorath, J.:. Interpretation of machine learning models using shapley values: application to compound potency and multi-target activity predictions. Journal of computer-aided molecular design **34**(10), 1013–1026 (2020) https://doi.org/10.1007/s10822-020-00314-0
9. Yang, H., Jiao, L. and Pan; Q.: A Survey on Interpretable Clustering, In: Proceedings of the 40th Chinese Control Conference (CCC), pp. 7384–7388. IEEE (2021) doi: 10.23919/CCC52363.2021.9549986.

Embedded Interpretable Regression using Dempster-Shafer Theory

Nelson Baloian[1][0000−0003−1608−6454], Edgar Davtyan[6][0009−0007−1356−2220], Karen Petrosyan[2][0009−0000−8038−7337], Arnak Poghosyan[3][0000−0002−6037−4851], Ashot Harutyunyan[4,5][0000−0003−2707−1039], and Sergio Penafiel[1][0000−0002−0025−7805]

[1] Department of Computer Science, University of Chile
{nbaloian,spenafie}@dcc.uchile.cl
[2] American University of Armenia karen_petrosyan2@edu.aua.am
[3] Institute of Mathematics NAS RA arnak@instmath.sci.am
[4] Institute for Informatics and Automation Problems NAS RA
[5] Yerevan State University harutyunyan.ashot@ysu.am
[6] Picsart edgar.davtyan@picsart.com

Abstract. This paper introduces an innovative approach to regression analysis by incorporating the Dempster-Shafer theory to enhance the interpretability and accuracy of regression models. Our method, Embedded Interpretable Regression (EI Regression), segments continuous output variables into discrete interpretable classes and applies targeted regression models to these classes. We demonstrate the potential and effectiveness of our approach through initial experimental validation, showing that our model achieves competitive accuracy compared to traditional regression methods while significantly improving the model's interpretability. This work contributes to interpretable machine learning and offers a practical framework for applying Dempster-Shafer's theory in predictive modeling.

Keywords: Embedded Interpretable Regression · Interpretable Machine Learning · Dempster-Shafer Theory · Uncertainty Modeling

1 Introduction

In the realm of machine learning, developing models that not only predict accurately, but also provide insights into their decision-making processes is critical. This necessity is particularly pronounced in fields where decisions have significant consequences, such as healthcare, financial forecasting, and legal assessments. Traditionally, there is a noticeable trade-off between the accuracy of a model and its interpretability. Highly accurate models, such as deep learning networks, often function as "black boxes," where the decision processes are obscure. Conversely, simpler models like decision trees or linear regressions offer clarity but usually at the cost of performance on complex tasks.

The growing demand for interpretable and accurate models has led to new research exploring methodologies that can bridge this gap. The objective is to

develop techniques that do not force practitioners to choose between understanding their model and achieving high predictive performance. Interpretable machine learning is crucial for complying with regulatory requirements, gaining user trust, and facilitating the adoption of AI systems in sensitive areas.

In this study, "embedded" refers to the integration of interpretability mechanisms directly within the regression model, rather than as a post-hoc analysis step. By embedding interpretability features within the model, our approach ensures that these features are an integral part of the learning process while maintaining predictive performance relative to the low or non-interpretable regression methods.

2 Related Work

Exploring the landscape of related work reveals a variety of approaches aimed at enhancing model interpretability without severely compromising accuracy:

Linear and logistic regressions have long been celebrated for their simplicity and interpretability. They provide clear insights into the relationships between independent variables and the predicted output but struggle with complex, non-linear relationships.

Decision trees and rule-induction systems excel in scenarios where it is crucial to understand the paths taken to reach a decision. They include C5.0 [10] and RIPPER [3] algorithms, which generate if-then rules, making them highly transparent but often less accurate with larger, more complex datasets.

Ensemble methods, including Random Forests [2] and Gradient Boosting Machines (GBMs) [4], deliver robust performance on diverse datasets. However, their ensemble nature, which involves aggregating predictions from numerous models, generally obscures the interpretability.

Bayesian methods offer a probabilistic perspective on modeling, providing insights into prediction uncertainty. However, they can become computationally intensive and less transparent with scale.

C5.0 algorithm is an extension of the decision tree concept, which builds more complex decision trees by creating rules that cover exceptions and special cases. C5.0 is known for their efficiency and the ability to handle missing data and large datasets effectively. RIPPER (Repeated Incremental Pruning to Produce Error Reduction): This rule-based learning algorithm constructs a set of rules from a training dataset and then prunes them to increase generalization. RIPPER is celebrated for its effectiveness in classification tasks and its ability to produce easily interpreted models.

Approaches like LIME (Local Interpretable Model-agnostic Explanations) [9] and SHAP (Shapley Additive Explanations) [6] have emerged as tools to explain predictions from black-box models. These methods attempt to provide post-hoc interpretability by approximating the local behavior of complex models with simpler, explainable models.

Innovations in Deep Learning: TabNet [1], a deep learning architecture, effectively blends interpretability with performance for tabular regression by em-

ploying a sequential, attention-based feature selection mechanism. Each decision step in TabNet involves selecting the most informative features through a trainable mask, enhancing instance-wise interpretability by elucidating the model's reasoning behind individual predictions. This approach allows TabNet to provide global insights through feature importance scores and instance-specific insights via attention masks. Such functionality makes TabNet highly competitive on various tabular regression tasks, outperforming traditional methods and offering more flexibility in modeling complex relationships. Unlike model-agnostic methods like LIME [9] and SHAP [6], which provide post-hoc explanations, TabNet integrates explainability directly into its architecture, enhancing its interpretative capabilities.

3 Contribution of this Work

This paper introduces a novel approach to regression analysis that leverages the Dempster-Shafer theory to create a framework that inherently balances interpretability with predictive accuracy. By segmenting the data into interpretable classes and applying targeted regression within these classes, the model aims to maintain transparency in how decisions are derived while striving to match or exceed the accuracy of traditional regression techniques.

This approach represents a significant step forward in the ongoing effort to develop machine learning models that practitioners can both trust and understand. The following sections will delve deeper into the theoretical underpinnings of this model, describe the experimental setup, and present a comprehensive analysis of its performance across several datasets.

4 Methodology

The embedded interpretable regression method seeks to deliver both high interpretability and accuracy. Traditional models are accurate but often fail to explain their predictions in understandable terms. This study introduces a regression framework utilizing the Dempster-Shafer theory (DST) to enhance both interpretability and accuracy [5, 7]. Interpretability is defined here as providing rules or guidelines that are understandable and verifiable, aiding users in comprehending the model's decision-making process.

By embedding interpretability mechanisms within the regression method, the approach ensures that interpretability features are an integral part of the learning process, thus supporting the method's predictions while maintaining its performance. The DST provides a flexible, robust framework for managing uncertainty, ideal for statistical models requiring evidence from varied sources. This approach enables informed predictions even with incomplete data.

The framework's core is the mass assignment function, defined as:

$$m : 2^X \rightarrow [0, 1], \tag{1}$$

where X represents all possible outcomes, and $m(A)$ the probability for each subset A of X. Belief and plausibility metrics support and potentially support a proposition, respectively:

$$\text{Bel}(A) = \sum_{B \subseteq A} m(B), \tag{2}$$

$$\text{Pl}(A) = \sum_{B \cap A \neq \emptyset} m(B). \tag{3}$$

The model integrates Dempster-Shafer's theory to structure the regression framework by defining a discerning frame and assigning probabilities to outcomes. During training, these probabilities are adjusted based on data insights.

Utilizing Dempster's rule of combination, the model integrates multiple evidence sources:

$$m_c(A) = \frac{1}{1-K} \sum_{B \cap C = A \neq \emptyset} m_1(B) \cdot m_2(C), \tag{4}$$

where K is a normalization factor, representing the degree of conflict between m_1 and m_2 given by:

$$K = \sum_{B \cap C = \emptyset} m_1(B) \cdot m_2(C)$$

The predictive model applies these rules to deliver highly interpretable decisions, illustrating clearly how predictions are derived from the evidence, supporting high interpretability and robust decision-making.

5 Proposed Algorithm

The experimental setup for the EI Regression approach is designed to optimize both interpretability and predictive accuracy by structurally decomposing the regression task into manageable segments. This process involves transforming a continuous output variable into a series of discrete categories, each represented as a unique bucket. This bucketization simplifies the complex regression problem, making the model's predictions and their bases more understandable [8]. Here are the detailed steps involved in setting up the experiment:

Bucketing: The first step in the EI Regression is to categorize the continuous dependent variable values into discrete classes or 'buckets'. This is done by dividing the range of the dependent variable into intervals. The bucketing technique employed, such as quantile bucketing, is crucial as it affects the distribution of data points across the buckets. Quantile bucketing, for example, divides the data into buckets such that each bucket contains approximately the same number of data points, thereby ensuring statistical significance and uniformity in each category. This step reduces variability within each bucket, which can significantly enhance the accuracy of subsequent predictions.

Classifier training: Once the data is bucketed, a classifier is trained to assign new instances to these predefined buckets. In the EI regression method, a

Dempster-Shafer classifier is utilized for this purpose [8]. This type of classifier is chosen for its ability to handle uncertainty and provide a measure of confidence in its classifications, which is derived from the Dempster-Shafer theory of evidence. The classifier uses mass assignment functions, which are probabilistic functions derived from the training data, to evaluate the likelihood of each new instance belonging to a particular bucket.

Regressor training: After classification, a specific regressor is trained for each bucket. This segmentation allows the model to tailor its regression algorithms to the particular characteristics of the data within each bucket, enhancing precision. For instance, simpler linear regression models might be used for buckets with linear relationships, while more complex models like gradient boosting or random forests could be employed for buckets with non-linear relationships. Please note that the choice of the regression model affects only the performance of the EI Regression method, the interpretability is assured by the DS Classifier. Therefore, in its essence, the EI Regression breaks down into the following 4 major steps:

1. **Data Bucketing:** Transform the continuous output variable into discrete buckets using techniques like quantile bucketing.
2. **Training Classifiers:** Train a Dempster-Shafer classifier to assign instances to buckets, leveraging DST to handle uncertainty.
3. **Training Regressors:** Within each bucket, train an appropriate regression model (linear regression, gradient boosting, or random forests) based on the data characteristics.
4. **Prediction:** For a new instance, classify it into a bucket, then apply the corresponding regressor to predict the output.

Figure 1 illustrates the complete process from training through prediction, showcasing how data is bucketed, classified, and then processed through specific regressors.

6 First Experiments and Conclusions

The experimental results demonstrate the effectiveness of the EI regression approach across various datasets. It achieves competitive accuracy compared to traditional regression models while significantly enhancing interpretability. For instance, in datasets with clearly definable categories, the model successfully predicts the bucket for new instances and the precise value within that bucket. This approach provides a clear interpretative advantage and elucidates the decision-making process through the rules derived from the Dempster-Shafer theory.

The bucketing method proved crucial in enhancing interpretability. Quantile bucketing was particularly effective as it distributed instances evenly, allowing for more stable and reliable regression within each bucket. The choice of regression algorithm for each bucket can be tailored based on the specific data characteristics within that bucket, offering a flexible approach to achieving high accuracy in predictions.

Fig. 1. Schematic overview of the Embedded Interpretable Regression (EI Regression).

Acknowledgments. This research was funded by ADVANCE Research Grants from the Foundation for Armenian Science and Technology. The authors are thankful to the anonymous reviewers for very constructive recommendations and comments, which helped improve the presentation of this material.

References

1. Arik, S.O., Pfister, T.: TabNet: Attentive Interpretable Tabular Learning (2020), arXiv:1908.07442
2. Breiman, L.: Random Forests. Machine Learning **45**(1), 5–32 (10 2001)
3. Cohen, W.W.: Fast Effective Rule Induction. In: Prieditis, A., Russell, S. (eds.) Machine Learning Proceedings 1995, pp. 115–123. Morgan Kaufmann (1995)
4. Friedman, J.H.: Greedy Function Approximation: A Gradient Boosting Machine. The Annals of Statistics **29**(5), 1189–1232 (2001)
5. Kim, B., Doshi-Velez, F.: Introduction to interpretable Machine Learning. Proceedings of the CVPR 2018 Tutorial on Interpretable Machine Learning for Computer Vision, Salt Lake City, UT, USA **18** (2018)
6. Lundberg, S., Lee, S.I.: A Unified Approach to Interpreting Model Predictions (2017), arxiv:1705.07874
7. Miller, T.: Explanation in Artificial Intelligence: Insights from the social sciences (2018), arXiv:1706.07269
8. Peñafiel, S., Baloian, N., Sanson, H., Pino, J.A.: Applying Dempster–Shafer theory for developing a flexible, accurate and interpretable classifier. Expert Systems with Applications **148**, 113262 (2020)
9. Ribeiro, M.T., Singh, S., Guestrin, C.: "Why Should I Trust You?": Explaining the Predictions of Any Classifier (2016), arxiv:1602.04938
10. Salzberg, S.L.: C4.5: Programs for Machine Learning by J. Ross Quinlan. Machine Learning **16**(3), 235–240 (1994)

Improving the DSGD Classifier with an Initialization Technique for Mass Assignment Functions

Aik G. Tarkhanyan[1][0009−0000−7015−111X] and Ashot N. Harutyunyan[2,3][0000−0003−2707−1039]

[1] Mathematics and Mechanics Department at Yerevan State University, 0025 Yerevan, Armenia
hayk.tarkhanyan@edu.ysu.am
[2] ML Lab at Yerevan State University, 0025 Yerevan, Armenia
[3] Institute for Informatics and Automation Problems NAS RA, 0014 Yerevan, Armenia
harutyunyan.ashot@ysu.am

Abstract. Several studies have shown that the Dempster–Shafer theory (DST) can be successfully applied to scenarios where model interpretability is essential. Although DST-based algorithms offer significant benefits, they do face challenges in terms of efficiency. We present a method for the Dempster-Shafer Gradient Descent (DSGD) algorithm that significantly reduces training time—by a factor of 2.1—and also reduces the uncertainty of each rule (a condition on features leading to a class label) by a factor of 1.4, while preserving accuracy comparable to other statistical classification techniques. Our main contribution is the introduction of a "confidence" level for each rule. Initially, we define the "representativeness" of a data point as the distance from its class's center. Afterward, each rule's *confidence* is calculated based on *representativeness* of data points it covers. This confidence is incorporated into the initialization of the corresponding Mass Assignment Function (MAF), providing a better starting point for the DSGD's optimizer and enabling faster, more effective convergence. The code is available at https://github.com/HaykTarkhanyan/DSGD-Enhanced.

Keywords: Dempster-Shafer Theory · Interpretability · Mass Assignment Functions · Classification.

1 Motivation

Dempster-Shafer theory [1] has emerged as a powerful framework for developing classification algorithms that prioritize interpretability. This theory provides a mathematical approach for combining evidence from different sources to calculate the probability of an event, utilizing Dempster's rule of combination. Peñafiel et al. [2] have demonstrated that algorithm combining Dempster-Shafer theory with optimization techniques can offer substantial explainability, even when employing a limited number of rules, without sacrificing accuracy. The algorithm is

inherently explainable because it operates by combining relatively simple rules for inference. To make improvements, two issues should be considered. Firstly, the number of subsets in the frame of discernment, growing with a complexity of $O(2^n)$, makes the inclusion of non-singleton classes nearly impossible. Secondly, given that each feature in the dataset typically generates three rules, combining even two rules significantly increases training time and further limits the predictive power of the rules. It becomes evident that to enhance DST-based models, a strategy for training time reduction is crucial.

2 Methodology

2.1 Overview

Our approach significantly reduces the optimization time by improving the technique for initializing Mass Assignment Functions associated with the rules. Peñafiel et al. [2] relied on random assignment, where the empty set's mass is set to 0, the entire set—which represents total uncertainty—is assigned a value of 0.8, and the remaining 0.2 is randomly distributed among singleton classes (other classes are not considered). We offer a technique that incorporates additional information about each rule into the Mass Assignment Function.

2.2 Representativeness Estimation

First of all, we need to define the representativeness of each data point. We calculate the Euclidean distance between each data point and the corresponding class center. We then apply outlier removal techniques and normalize the data via Min-Max scaling. The resulting value is used as a representativeness measure.

2.3 Rule Confidence Estimation

Utilizing the algorithm previously described, we calculate the representativeness of each data point. Afterward, we generalize this data point-specific estimate to an entire rule using the following steps:

1. Filter the dataset to retain only the rows that comply with the rule.
2. If the rule does not apply to any rows, set its confidence to 0. Otherwise:
3. Calculate the rule's confidence as the mean representativeness of the rows it covers.
4. If the rows are not homogeneous with respect to their labels, reduce the confidence based on the proportion of the most frequent label among these rows.

2.4 Mass Assignment Function Initialization

After confidence calculation, we initialize the MAF. Let $c = confidence(rule)$ represent the confidence derived for a given rule. The label l_{mode}, which is the most frequently occurring label within the subset of data points covered by the rule, receives the confidence value c. The remaining mass, $(1 - c)$, is evenly distributed among the rest of the labels present in the subset. Formally, for an element l_i in the subset:

$$m(l_i) = \begin{cases} c & \text{if } l_i = l_{\text{mode}}, \\ \frac{1-c}{n-1} & \text{otherwise,} \end{cases}$$

where $m(l_i)$ denotes the mass assigned to label l_i, and n is the total number of elements in the frame of discernment.

3 Results

Here we demonstrate the effects of the newly defined MAF initialization algorithm on the training time, accuracy, and the amount of rule uncertainties. We accomplish this by testing the approach both on controlled scenarios and on some classical datasets. The datasets used for evaluations are summarized in Table 1.

Table 1: Datasets Overview (Binary Classification)

Dataset	**Rows**	**Columns**	**Description**
Brain Tumor	3762	14	Includes first-order and texture features with target levels.
Breast Cancer Wisconsin	699	9	Clinical reports detailing cell benignity or malignancy.
Gaussian	500	3	Two 2D Gaussian distributions generate this dataset.
Uniform	500	3	Uniform samples from [-5, 5], with class split by the sign of x.
Rectangle	1263	3	Points in [-1, 1]×[-1, 1], class determined by the y component's sign.

The first two ([3,4]) are real-life datasets, while the last 3 are controlled scenarios.

3.1 Accuracy and Speedup Analysis

Moving forward, we will refer to our MAF initialization technique as "Confidence", the median as "MED." and the average as "AVG.".

Table 2 presents a comparison of various metrics across different MAF initialization methods (Confidence and Random) and datasets. For evaluating the classifier's predictive power, we have calculated the accuracy (ratio of correctly predicted instances to the total instances) and F1 score (harmonic mean of the precision and recall). For evaluating the optimizer, we have reported the training time in seconds, the number of epochs, the minimum loss, and the initial loss.

Table 2. Comparison of Various Metrics Across Different MAF Initializations and Datasets

MAF Method	Dataset	Accuracy	F1	Training Time, s	Epochs	Min Loss	Initial Loss
confidence	Brain Tumor	0.981	0.979	72.832	75	0.023	0.410
random	Brain Tumor	0.983	0.981	288.722	133	0.026	0.241
confidence	Breast Cancer	0.966	0.952	32.671	82	0.022	0.614
random	Breast Cancer	0.971	0.959	31.940	111	0.030	0.336
confidence	Gaussian	0.987	0.988	41.419	135	0.020	0.148
random	Gaussian	0.967	0.969	91.876	273	0.023	0.282
confidence	Rectangle	1	1	122.795	168	0.006	0.165
random	Rectangle	1	1	212.407	287	0.007	0.253
confidence	Uniform	0.973	0.970	56.482	191	0.037	0.198
random	Uniform	0.973	0.970	82.152	273	0.038	0.254

We can see that our method requires fewer epochs to converge, and it converges to a smaller loss. Table 3 further examines the results by comparing the training time speedup and the ratios of accuracy, F1 score, and for the Confidence and Random MAF initialization methods across different datasets.

Table 3. Ratios of Accuracy, F1 Score, and Training Time for Random and Confidence MAF Initializations Across Datasets

Dataset	Accuracy Ratio	F1 Ratio	Training Time Speedup, x
Brain Tumor	1.00	1.00	3.96
Breast Cancer	0.99	0.99	0.98
Gaussian	1.02	1.02	2.22
Rectangle	1.00	1.00	1.73
Uniform	1.00	1.00	1.45

3.2 Uncertainty Analysis

Now we'll take a look at the effect of our new MAF initialization on the uncertainty levels of the rules. To do this, we compare the average and median uncertainties for both the random initialization and the confidence-based initialization of the MAF. The results are summarized in Table 4.

The improvement factor is defined as the ratio of $MED\ Rand.$ and $MED\ Conf.$.

Table 4. Average and Median Uncertainties for Random and Confidence MAF Initializations

	AVG Rand.	AVG Conf.	MED Rand.	MED Conf.	Improvement
Brain Tumor	0.71	0.42	0.74	0.42	1.76
Breast Cancer	0.72	0.46	0.71	0.47	1.53
Gaussian	0.33	0.27	0.31	0.27	1.15
Rectangle	0.30	0.27	0.34	0.24	1.42
Uniform	0.28	0.19	0.11	0.11	1.04

On average, the confidence approach yields a reduction in uncertainty by a factor of 1.38.

4 Conclusion

In conclusion, this work presents a novel approach to initializing Mass Assignment Functions in Dempster-Shafer Theory. By introducing a "confidence" level for each rule based on the "representativeness" of the data points it covers, we have demonstrated a significant reduction in training time—by an average factor of 2.06—without reduction in accuracy or F1 scores. Additionally, on average our method reduces rule uncertainty by a factor of 1.38.

We evaluated our approach on both real-life and controlled datasets and showed that proposed MAF initialization method leads to faster convergence, fewer epochs, and lower minimum loss compared to random initialization. In particular, the reduction in training time makes it more feasible to include non-singleton classes in DST-based classifiers and allows us to add more rules to the algorithm.

Additionally, we have made the codebase available at https://github.com/HaykTarkhanyan/DSGD-Enhanced

4.1 Future Work

In this work, the confidence of the rule used in MAF initialization has been calculated as a mean of Euclidean distances from the corresponding class center (simple mean of the points belonging to the class). Using KMeans to estimate the class center can lead to improvements.

Additionally, a drawback of our approach emerges when the data does not have a spherical shape. In such cases, the Euclidean distance can be an ineffective measure. We plan to explore approaches which are based on the density of the data.

Furthermore, the testing of our approach has been done only on datasets containing numeric features. We plan to explore approaches for datasets that contain categorical variables.

Acknowledgement. The research was supported by ADVANCE Research Grants from the Foundation for Armenian Science and Technology.

References

1. Shafer, G.: A Mathematical Theory of Evidence. Princeton University Press, Princeton (1976)
2. Peñafiel, S., Baloian, N., Sanson, H., & Pino, J. A.: Applying Dempster–Shafer theory for developing a flexible, accurate and interpretable classifier. Expert Systems with Applications **148**, 113262 (2020)
3. Wolberg, W. H., Mangasarian, O. L.: Multisurface method of pattern separation for medical diagnosis applied to breast cytology. Proceedings of the National Academy of Sciences **87**(23), 9193–9196 (1990)
4. Bohaju, J.: Brain Tumor. In: Kaggle 2020, DOI: https://doi.org/10.34740/KAGGLE/DSV/1370629. https://www.kaggle.com/dsv/1370629 (2020)

An Empirical Analysis of Feature Engineering for Dempster-Shafer Classifier as a Rule Validator

Aneta Baloyan[2], Alexander Aramyan[3], Nelson Baloian[1][0000−0003−1608−6454], Arnak Poghosyan[4][0000−0002−6037−4851], Ashot Harutyunyan[5,6][0000−0003−2707−1039], and Sergio Penafiel[1][0000−0002−0025−7805]

[1] Department of Computer Science, University of Chile
nbaloian,spenafie}@dcc.uchile.cl
[2] Metric, Armenia aneta.baloyan@gmail.com
[3] American University of Armenia alex.aramyan@proton.me
[4] Institute of Mathematics NAS RA arnak@instmath.sci.am
[5] Institute for Informatics and Automation Problems NAS RA
[6] Yerevan State University harutyunyan.ashot@ysu.am

Abstract. Explainable AI methods are increasingly attracting the practitioner's attention since they convey important information about the nature of the phenomena being studied. Rule-generating models are considered one of the most explainable ones, however, there are so far not many attempts aimed at evaluating the rules generated by this kind of models. This paper proposes using an explainable classifier based on the Dempster-Shafer (DS) plausibility theory as a rule-validating mechanism to assess the reliability of the rule sets generated by AI models. The DS theory enables combining evidence from various sources while dealing with conflicting or even contradictory information, identifying trustworthy rules with high belief correctness. Our empirical analysis evaluates the DS-based classifier's ability to learn possibly complex numeric feature interactions on synthetic datasets. Results show the model excels at numeric interactions like differences and ratios and performs well regardless of class imbalance, but struggles with imbalanced categorical data. We conclude by proposing future work including comparative result analysis against traditional methods and developing hybrid approaches for more robust but at the same time interpretable AI systems.

Keywords: Explainable AI · Dempster-Shafer theory · Rule extraction · Rule validation · Feature interactions · Interpretable machine learning.

1 Introduction

The rise of artificial intelligence (AI) has revolutionized various industries, enabling us to tackle complex problems with unprecedented efficiency and effectiveness. However, as AI models grow in complexity, their inner workings become increasingly opaque, making it challenging to understand and validate the reasoning behind their decisions. Explainable AI (XAI) [1] has emerged as a crucial field, aiming to shed light on these black-box models and enhance trust in their

outputs. Among the various approaches to XAI, rule extraction techniques hold significant promise [2]. By analyzing trained black-box models and extracting interpretable rules that approximate their decision-making logic, rule extraction offers a bridge between complex models and human-understandable explanations. However, the extracted rule sets may exhibit inconsistencies, conflicts, or unreliable patterns, hampering their effectiveness in enhancing model interpretability. We propose to use an existing Dempster-Shafer (DS) classifier [3] as a rule validation mechanism to assess and improve the reliability of extracted rule sets. The DS theory, a powerful framework for reasoning with uncertain and imprecise information, provides a principled approach to combining evidence from multiple sources while accounting for potential conflicts or complementary evidence. For example, in the context of stroke prediction in healthcare [4], the DS classifier can combine belief functions from extracted rules suggesting high or low risk, accounting for agreements and conflicts, to provide a comprehensive assessment of the rule set's reliability. Rules with high belief in their correctness and low conflict with other rules can be identified as more trustworthy, while rules with low reliability or high conflicts can be flagged for further investigation or refinement. Moreover, the DS classifier's ability to incorporate domain knowledge or expert opinions into the validation process enhances its applicability in real-world scenarios. Initial belief functions can be assigned based on prior knowledge or expert assessments, further refining the validation process and aligning it with domain-specific requirements. Leveraging the DS classifier for rule validation would improve the interpretability and trustworthiness of AI systems, particularly in critical domains such as healthcare, where transparency and accountability are paramount. In this work, we present an empirical analysis of the DS classifier's feature interaction learning capabilities, its strengths and weaknesses, the extent of provided explanability for common use cases, suggesting its credibility for being utilized as a rule validating tool. This work focuses on the ability of the instrument to detect certain mathematical relations between the attributes of a set of samples. While our study focuses on synthetic datasets, the insights gained are highly relevant to real-world applications. The carefully designed feature interactions in our synthetic data - including counts, differences, logarithms, powers, ratios, and products - mimic the types of complex relationships often found in real-life datasets across various domains. By systematically evaluating the DS classifier's performance on these controlled scenarios, we can assess its capacity to capture and learn intricate patterns that frequently occur in real-world data. The use of synthetic data allows us to isolate specific types of interactions and precisely measure the model's learning capabilities. This controlled environment provides a clear understanding of the DS classifier's strengths and limitations, which can then inform its application to real-world problems. For instance, the model's demonstrated proficiency in learning nonlinear numeric interactions suggests it could be particularly effective in fields like finance or physics, where such relationships are common. Moreover, by varying factors like dataset size and class balance, we simulate challenges often encountered in real-world data analysis. This approach helps us understand how the DS

classifier might perform under different real-life conditions, providing valuable insights for practitioners considering its use in various applications.

We do not present here a comparison of the DS Classifier model performance with other classifiers' performance, as this has already been studied and reported in the original paper presenting the Dempster-Shafer classifier for the first time [3]. The paper is organized as follows, section 2 briefly lists the prior work done to this point, the DS theory, and other relevant studies, section 3 presents the empirical analysis design in detail, section 4 discusses the initial outcomes of the analysis. Section 5 concludes and discusses directions for future work.

2 Related Work

The DS classifier is a unique machine learning model that differs from traditional approaches in its underlying principles. Unlike probabilistic classifiers that rely on explicit probability distributions, the DS classifier operates on the concept of belief functions, which provide a more flexible and comprehensive representation of uncertainty. This approach enables the DS classifier to capture complex relationships between features, including non-linear interactions and synergies, in a transparent manner.

Previous research, such as the work presented by Peñafiel et al. [3], has delved into the theoretical foundations and practical applications of the DS classifier. These studies have highlighted the model's ability to handle incomplete or imprecise data, as well as its potential for decision-making in the presence of uncertainty.

Although the DS Classifier is rooted in the Dempster-Shafer theory, it is finally aimed at being used to solve classification problems, and according to this, its features should be evaluated using the for classifying Beyond the DS classifier, there exist other rule-based models that aim to provide explanations for their predictions. Decision trees, for example, generate interpretable if-then rules that can be easily understood by domain experts. The RIPPER [6] algorithm and Skope-rules [7] are additional examples of rule-based approaches that have been developed to enhance the transparency of machine learning models. However, each of these models, as well as other Machine Learning models, have their intrinsic limitations in terms of capturing feature interactions in the dataset. [8] highlights the strengths and weaknesses of most used Machine Learning models empirically analyzing their feature interaction learning capabilities.

3 Methodology

To empirically analyze the DS classifier's ability to learn complex feature interactions, we have adopted a systematic data generation approach. This approach involves creating random numeric features and calculating target variables based on specific mathematical expressions, which incorporate a range of operations, such as sums, products, ratios, exponents, polynomials, and counts.

The generated datasets are designed to mimic the types of feature interactions typically encountered in real-world machine learning tasks. To transform these datasets into classification problems, we have expressed the outcomes in terms of equalities and inequalities, where the target variable (Y) is assigned a value of 1 if the specified mathematical expression is satisfied and 0 otherwise.

We have considered two examples to illustrate the different levels of complexity in the data:

Table 1: Example 1.

X_1	X_2	$Y = (X_1 + X_2 > 1)$
-23.1	21	0
1	3	1

The example shown in Table 1 represents a static complexity, where the target is determined based on a fixed threshold. This simplicity allows for a clear understanding of how the DS classifier handles basic non-linear relationships.

Table 2: Example 2.

X_1	X_2	X_3	$Y = (X_1 + X_2 > X_3)$
-23.1	21	1	0
1	3	5	0
1	3	2	1

The example shown in table 2 introduces dynamic complexity by varying the threshold X_3, adding an extra layer of interaction. This scenario more closely resembles real-world situations where feature interactions may not only be non-linear but also variable.

In this study, we have chosen to focus on Example 1 for several reasons. The primary objective is to establish a foundational understanding of the DS classifier's ability to learn complex feature interactions. The simpler scenario in Example 1 provides the necessary clarity and control to isolate the effects of the classifier's performance, without the added complexities of a variable threshold.

Moreover, the constant threshold in Example 1 ensures a uniform environment for comparative analysis, where the DS classifier's performance can be benchmarked against traditional machine learning methods, such as logistic regression and decision trees. This consistency is crucial for a fair and effective comparison.

To evaluate the DS classifier's performance, we will train the model on the generated datasets and assess its ability to learn the underlying feature interactions. The model's performance will be measured using standard evaluation metrics, such as accuracy, F1 score, and ROC AUC based on the belief values

of the resulting Dempster mass function. We will consider the DS classifier to have successfully learned a complex interaction if it achieves 100% performance on the test dataset.

3.1 Data Generation

To empirically evaluate the DS classifier's ability to learn complex feature interactions, we generate synthetic datasets. Specifically, we create random numeric features and calculate target variables based on predetermined mathematical expressions:

1. **Counts:** Generate varying amounts of binary random variables. Calculate the numeric target as the number of variables that are equal to 1. Convert the numeric target into binary classes based on a simple threshold calculation, such as $target > t$, where t is a predetermined constant threshold.
2. **Differences:** Generate two numeric random variables. Calculate the numeric target as their difference. Convert the numeric target into binary classes based on a simple threshold calculation, such as $target > t$, where t is a predetermined constant threshold.
3. **Logarithms and Powers:** Generate one random variable. Calculate the numeric target as the logarithm/square/cube of the random variable. Convert the numeric target into binary classes based on a simple threshold calculation, such as $target > t$, where t is a predetermined constant threshold.
4. **Ratio and Product:** Generate two numeric random variables. Calculate the numeric target as their ratio/product. Convert the numeric target into binary classes based on a simple threshold calculation, such as $target > t$, where t is a predetermined constant threshold.

For each of these examples, we generate datasets of varying sizes (100, 1,000, and 10,000 samples) to investigate the model's performance and scalability. The target variables are expressed as binary classifications, where the model aims to predict whether the specified mathematical expression is satisfied or not.

3.2 Model Training and Evaluation

The DS classifier is trained on the generated datasets using statistical single rules. The rule generation method was chosen the same for all the datasets. To make the rules, and hence the learning capability of the model, utmost independent of the number of breaks, we implemented a "soft breaks" technique, where the model was not initiated with a single mutually exclusive set of rules, but all rulesets generated by 5 or fewer breaks have been added. The model's performance is evaluated using standard metrics such as accuracy, F1 score, and ROC AUC. To assess the model's ability to learn complex interactions, we consider it successful if it achieves 100% performance on the test dataset, as the data is generated to be precisely calculable.

The evaluation is performed on unseen test datasets, which are held out from the training process. This approach allows us to assess the model's generalization capabilities and its ability to capture the underlying feature interactions effectively. Additionally, the randomization of the dataset is calibrated in a way that lowers the chances of repetitive observations.

4 Results

The model was able to learn the count of features with 100% accuracy for the perfectly balanced datasets where the distribution of counts was close to normal. However, any slight imbalance in the dataset resulted in a drastic drop/decrease in model performance.

To evaluate the DS classifier's performance, we use the Receiver Operating Characteristic Area Under the Curve (ROC AUC) as a key metric. ROC AUC provides a holistic measure of the classifier's performance across all possible classification thresholds, comparing it against random choice. This metric is particularly valuable for assessing model performance on imbalanced datasets and allows for meaningful comparisons between different models. Many similar studies in the field also utilize ROC AUC as a performance metric, enabling us to benchmark our results against existing research. By using ROC AUC, we can comprehensively evaluate the DS classifier's ability to learn and generalize complex feature interactions.

As observed in the example shown in fig. 1, with increase in class imbalance from left to right, ROC AUC drops for the corresponding model. Note that the distribution of the target variable before applying the threshold criteria is close to normal in all three cases.

For the rest of the interactions, however, the class imbalance did not affect the model performance as much as in case of counts. As we can see from the plots in fig. 2, the model performed very well (with an F-1 score = 0.99 on an unseen test set) learning feature differences on an imbalanced dataset.

The model's performance on balanced datasets is invariant of the target's distribution. This is observed when evaluating power, logarithm, as well as ratio and product feature interactions, as seen in the fig. 3, squaring (3a), cubing (3b) and taking the logarithm (3c) of the feature does not affect the model performance. The same holds for feature ratios (3d) and products (3e). Hence, we can claim that the model's accuracy holds also for non-normal distributions.

4.1 Analysis

The results demonstrate the Dempster-Shafer classifier's strengths and weaknesses in learning various types of feature interactions. The model exhibited high performance in learning feature interactions on numeric columns (differences, squares, cubes, logarithms, ratios, and products), achieving near-perfect accuracy even with imbalanced datasets. However, the model's performance suffered significantly when learning to count the number of binary features relying on categorical variables only.

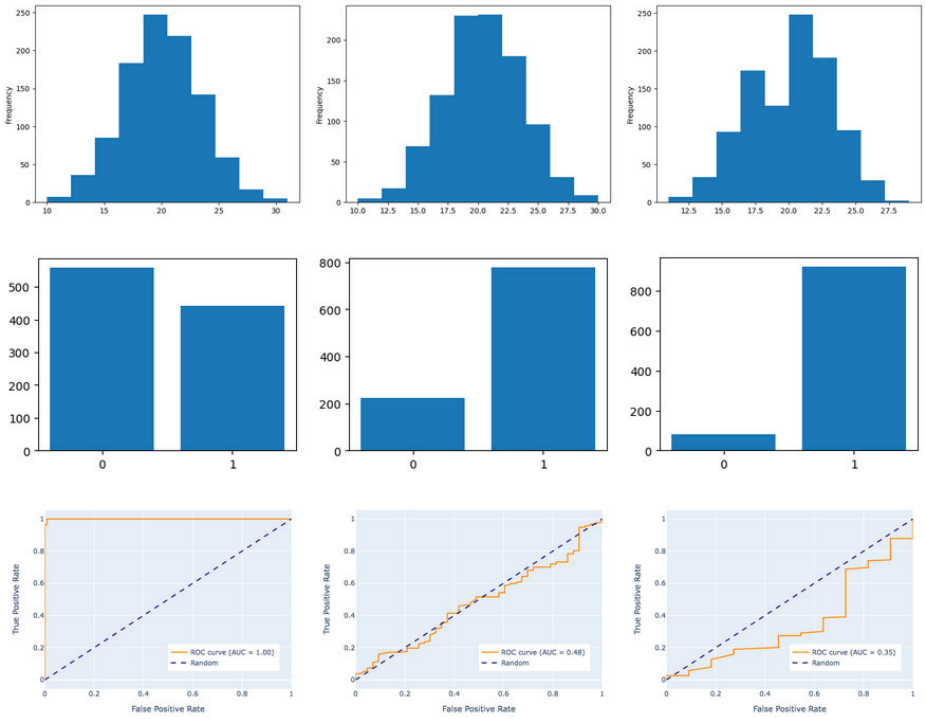


Fig. 1: The histograms of the target variables for "count" interaction before applying the threshold criteria are shown on the first row for each example dataset, the bar-plots on the second row show the corresponding class distribution after applying the threshold criteria, and the ROC curves are shown on the last row.

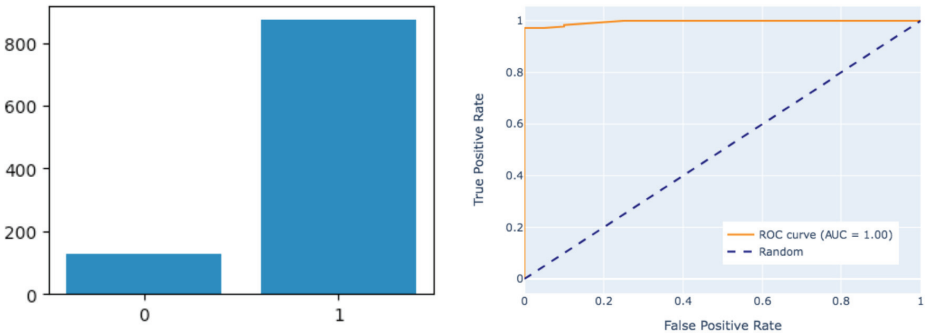
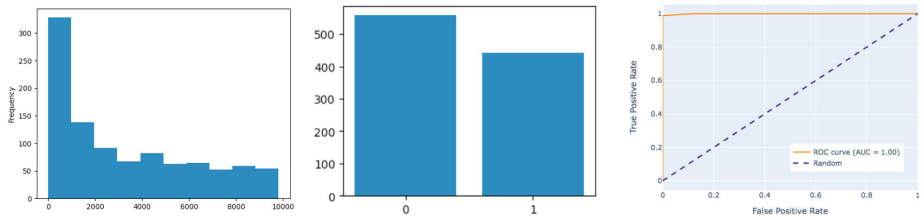
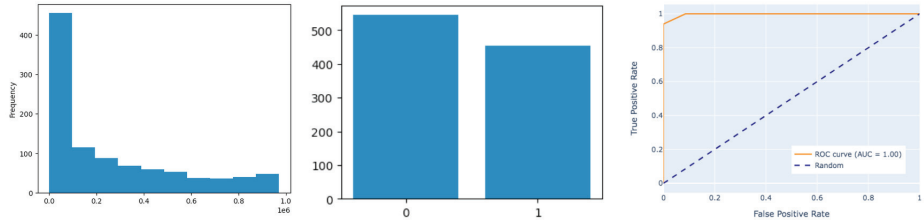


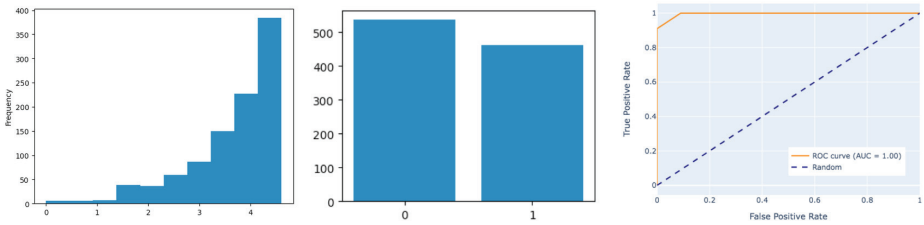
Fig. 2: The bar-plot shows the distribution of target classes after thresholding for "difference" feature interaction. ROC curve is shown on the right.



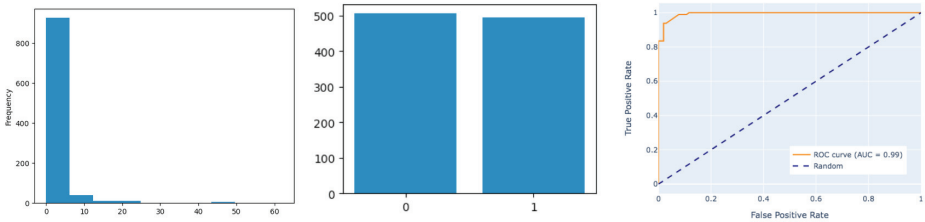
(a) square



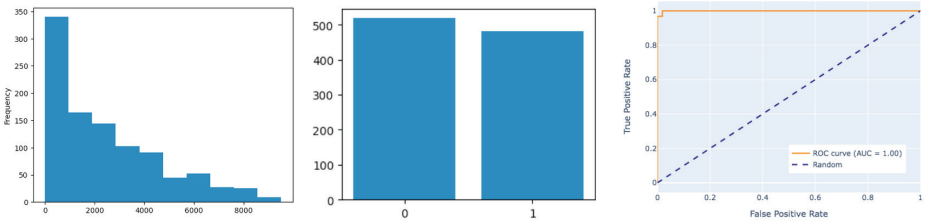
(b) cube



(c) logarithm



(d) ratio



(e) product

Fig. 3: Each subfigure contains the distribution of the target variable before thresholding, the class distribution after thresholding and the ROC curve in the specified order.

The model was also observed to lose accuracy with increasing amounts of class imbalance. This sensitivity to imbalanced data distributions highlights a potential limitation of the DS classifier in certain scenarios.

Furthermore, the choice of statistical breaks and their impact on the model's performance warrant further investigation. While the "soft breaks" approach aimed to mitigate the influence of the number of breaks, a more systematic analysis of this aspect could provide valuable insights into optimizing the DS classifier's rule generation process.

Another area for future exploration is the model's sensitivity to the target variable's distribution. While the generated datasets were designed to have specific mathematical expressions, understanding how the DS classifier handles different target distributions could broaden its applicability in real-world scenarios with varying data characteristics.

While our study focuses on synthetic datasets, the insights gained are highly relevant to real-world applications. The carefully designed feature interactions in our synthetic data - including counts, differences, logarithms, powers, ratios, and products - mimic the types of complex relationships often found in real-life datasets across various domains. By systematically evaluating the DS classifier's performance on these controlled scenarios, we can assess its capacity to capture and learn intricate patterns that frequently occur in real-world data. The use of synthetic data allows us to isolate specific types of interactions and precisely measure the model's learning capabilities. This controlled environment provides a clear understanding of the DS classifier's strengths and limitations, which can then inform its application to real-world problems. For instance, the model's demonstrated proficiency in learning non-linear numeric interactions suggests it could be particularly effective in fields like finance or physics, where such relationships are common. Moreover, by varying factors like dataset size and class balance, we simulate challenges often encountered in real-world data analysis. This approach helps us understand how the DS classifier might perform under different real-life conditions, providing valuable insights for practitioners considering its use in various applications.

The observed weaknesses of the DS classifier suggest potential areas for improvement or the development of hybrid approaches that combine its strengths with those of other machine learning models. By addressing these limitations, the DS classifier's ability to capture complex feature interactions could be further enhanced, leading to more robust and interpretable models for a wider range of applications.

5 Conclusion and Future work

The Dempster-Shafer (DS) classifier offers compelling advantages for rule discovery and validation in explainable AI. Its ability to effectively learn complex numeric feature interactions like differences, powers, logarithms, ratios, and products showcases potential for uncovering intricate non-linear patterns within real-world datasets. Furthermore, the DS classifier's grounding in Dempster-

Shafer theory enables principled reasoning under uncertainty. By combining evidence sources while accounting for conflicts or complementarity, it can assess the reliability of extracted rules - a crucial capability for enhancing AI system interpretability and accountability, especially in critical domains demanding transparency and well-founded explanations.

In the future work, we aim to provide the full results of the suggested analysis. We aim to conduct the analysis on a larger set of datasets with more randomization and statistically validate the results presented in this work. We also aim to add further mathematical expressions such as max of input, polynomials, distance, etc.

While the current study focuses on evaluating the DS classifier performance in learning complex feature interactions, we aim to conduct a comparative analysis with traditional machine learning methods, such as logistic regression and decision trees. By applying these baseline models to the same set of generated datasets, we can identify the relative strengths and weaknesses of the DS classifier in capturing various types of feature interactions.

The comparative analysis will delve into a detailed exploration of the specific interactions that each model excels at or struggles with. This future work will enable us to contextualize the DS classifier's capabilities within the broader landscape of machine learning models, highlighting its unique advantages and potential areas for improvement. Furthermore, the findings from the comparative analysis may inform the development of hybrid approaches that combine the strengths of different models, ultimately leading to more robust and interpretable solutions for handling complex feature interactions.

References

1. Xu, F., Uszkoreit, H., Du, Y., Fan, W., Zhao, D., Zhu, J. Explainable AI: A brief survey on history, research areas, approaches and challenges. In 8th CCF International Conference, NLPCC October 9–14, Proceedings, Part II 8, pp. 563-574. Springer International Dunhuang, China (2019).
2. Obregon, J., Jung, J. Y. . RuleCOSI+: Rule extraction for interpreting classification tree ensembles. Information Fusion **89** (2023) pp. 355-381.
3. Peñafiel, S., Baloian, N., Sanson, H., Pino, J. A. :Applying Dempster–Shafer theory for developing a flexible, accurate and interpretable classifier. Expert Systems with Applications **148** (2020) pp. 113262.
4. Penafiel, S., Baloian, N., Sanson, H., & Pino, J. A.: Predicting stroke risk with an interpretable classifier. IEEE Access, **9** (2020), pp. 1154-1166.
5. Chaki, J. . The Dempster–Shafer Theory to Handle Uncertainty in Artificial Intelligence. In Handling Uncertainty in Artificial Intelligence (pp. 25-35). Singapore: Springer Nature Singapore (2023).
6. Cohen, W. W. Fast Effective Rule Induction In Proceedings of the Twelfth International Conference on Machine Learning (1995) pp. 115-123.
7. Goix, N. et al: ML learning with logical rules in Python; scikit-learn-contrib/skope-rules v1.0.1, December 2020. https://zenodo.org/records/4316671
8. Heaton, J.: An empirical analysis of feature engineering for predictive modeling. SoutheastCon 2016. https://doi.org/10.1109/secon.2016.7506650

Interpretability of Machine Learning Models in the Insurance Sector

Anna Sargsyan[1]

[1]PLAT.AI, Yerevan, Armenia
sargsyan.anna.g@gmail.com

Abstract. Recently, machine learning models become increasingly complex, and their interpretability becomes a critical concern. This paper investigates the challenges and methodologies associated with enhancing the interpretability of ML models, with a specific focus on their application within the insurance domain. It considers the fact that there are different stakeholders that need to understand how the model operates, and their needs can be covered with different approaches.

Keywords: Machine Learning, Insurance Sector, Interpretability, Black-Box Algorithm, Model Interpretability, Real Life Insights, Risk Score

1 Introduction

The financial services industry is undergoing a rapid and disruptive digital and AI transformation, and the insurance sector is not an exclusion. In insurance, AI/ML has significant potential to help reduce protection gaps by improving the availability, affordability, and accessibility of insurance on the back of increased personalization and improved cost-efficiency. In insurance, AI is most commonly used in underwriting, claims processing, customer service and fraud detection [1]. Various studies indicate that complex models generally outperform Generalized Linear Models (GLMs) in terms of predictive accuracy [2]. However, while linear models provide a straightforward interpretation of their predictions, Black-Box models lack a universally applicable solution for interpretation. Moreover, different stakeholders within the insurance domain necessitate varying levels of interpretability, prompting ML Engineers to address all their distinct needs [3]. Consequently, this paper concentrates on delineating the requirements of each stakeholder and proposing the most suitable interpretability technique for each.

2 Research Methodology

In this research, practical application of available methodologies is emphasized using different perspectives of stakeholders in the insurance domain. Real-life data from the Armenian car insurance market is used. However, in order not to reveal any commercial secret, a subset of data and subset of significant variables was used for training models and interpreting results. The models trained are targeting the probability that the policyholder will experience a claim in the requested policy period, hence binary classification methods are applied.

Afterward, the predicted probabilities are turned into Risk Scores, which are used by actuaries to determine prices for each police. Both linear and Black-Box complex models

were trained and examined, however the CatBoost[1], which showed the highest accuracy metrics among tested models, was chosen to illustrate the analysis. Different interpretability techniques were applied, including Feature Importance Plots, LIME, surrogate models, and SHAP values. Among these, SHAP values provided the most comprehensive and intuitive explainability. In addition, global and individual SHAP values cover most of the stakeholder needs, which is why they are emphasized below [4], [5].

3 Findings and Discussion

3.1 Model Explainability Using SHAP Values: Application to Real-Life Data

SHAP values provide a structured and consistent way to interpret the decisions of the CatBoost model by highlighting feature importance, effect, interactions, and providing instance-level insights. The dots in the SHAP values summary plots visualize the SHAP values for the records in the training subsample. The color of a dot represents the feature value for the specific data point, and the position on the x-axis displays the corresponding SHAP value. A high SHAP value indicates a higher (positive) predicted probability by the model, whereas a lower (negative) SHAP value contributes negatively to the predicted probability by the model [6]. When examining Global SHAP values calculated for all risk groups (Fig. 1.) it becomes apparent that higher values for the "Historical risk score" feature typically align with elevated SHAP values, thereby indicating an increased likelihood of the policyholder filing a claim. Conversely, a high value for the "Duration insured" feature generally suggests a lower probability of a policyholder making a claim. The SHAP values for the "Number of non-guilty accidents" and "Vehicle hp" variables exhibit both positive and negative impacts, indicating a substantial variation in its influence on predictions across the dataset.

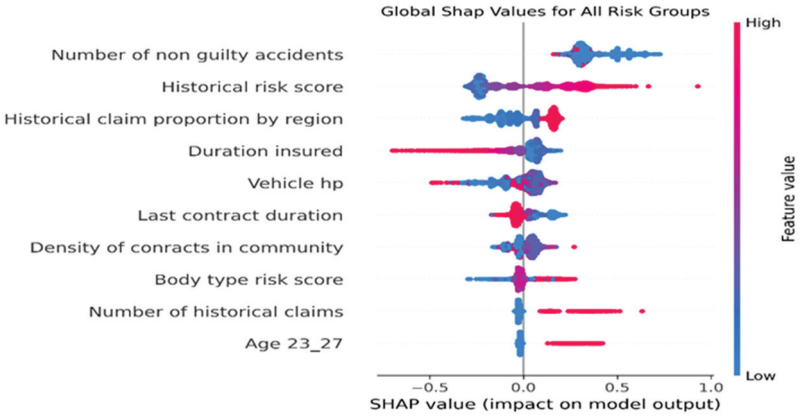


Fig. 1. Global SHAP values calculated for all risk groups.

For investigating the influence of various factors on different risk categories (scores) of policyholders, SHAP values were computed and compared for two distinct groups.

[1] https://catboost.ai/

The first group comprises policyholders with a risk score of 1, while the second includes those with risk scores of 9 and 10. An examination of Fig. 2. and Fig. 3. revealed disparities of the ranking and directional impact of influential variables between these groups, which provides quite deep intuition and information to cover different stakeholder needs.

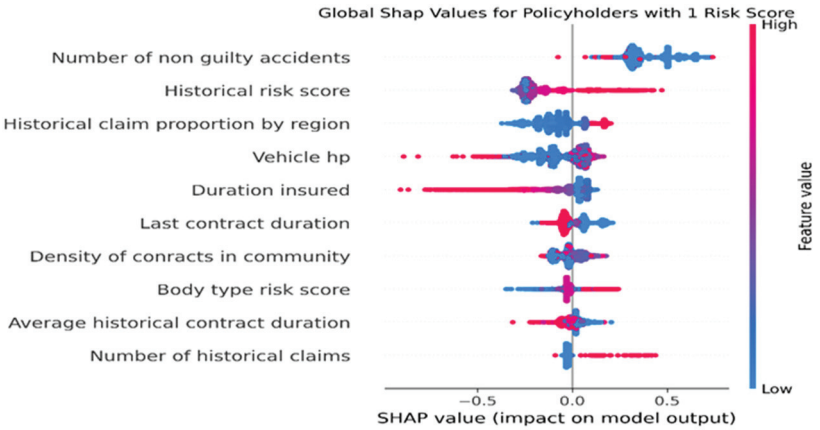


Fig. 2. Global SHAP values calculated for policyholders with 1 risk score.

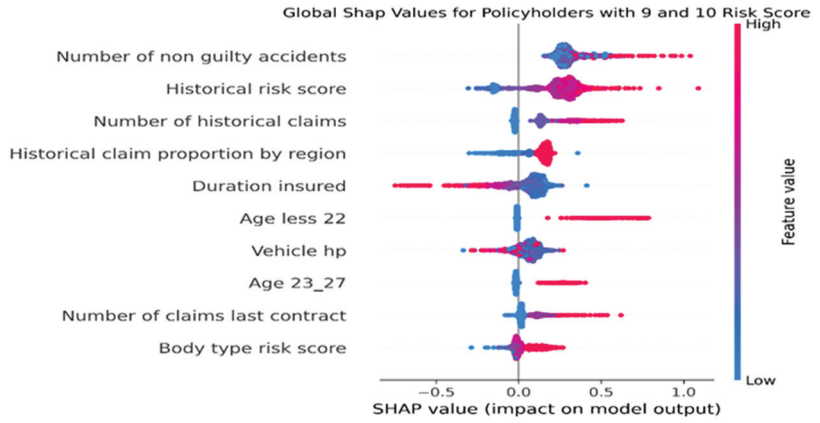


Fig. 3. Global SHAP values calculated for policyholders with 9 and 10 risk score.

3.2 Key Stakeholders of Risk Models in the Insurance Space and Analysis of Their Needs

Regulators

Insurance regulators in all countries oversee the sector, ensuring consumer rights with transparent and fair policies while enforcing rules to prevent anti-competitive behavior. They might prohibit utilization of some variables, which implies discrimination. They are also responsible for making sure that all players have the same access to the govern-

mental databases. The regulators in some countries impose a rule that the insurance company should be able to provide reasoning behind the specific pricing of the insurance policy if the policyholder requests it [7]. Depending on the desired depth of model interpretability, various approaches can be employed. These range from simply listing variables to utilizing Feature Importance Plots, LIME, and conducting more detailed analyzes using Global and Local SHAP values. In case of regulatory implication of providing explanations to the policyholders directly, individual SHAP values are identified as the most suitable method using the literature and own practical research as they appeared to be the most intuitive and accurate.

Actuaries

When insurance companies use ML for underwriting, actuaries base their pricing strategies on model predictions, focusing on accident frequency and severity. They leverage their expertise to deeply analyze and validate model performance, considering elimination of model's focus on temporary trends and emphasizing potentially significant factors for future accuracy. Actuaries also assess how their models and pricing decisions impact market dynamics and customer segmentation, recognizing that pricing significantly influences consumer choices. They scrutinize different segments of predicted probabilities and key features at each score level to identify potential inaccuracies using test data [8]. Hence, the actuaries need to investigate the lists of all variables used, the probability distribution of the predictions, coefficients of linear models, Feature Importance Plots, Global and Local SHAP values and their distribution across different score segments.

Sales and Marketing Team

While traditional sales approaches focus on maximizing customer acquisition, many respected insurance companies adopt a selective strategy. To build a profile of an ideal customer, strategists use a list of variables that influence accident risk. This enables them to identify controllable features for targeted marketing campaigns, utilizing Global SHAP values to guide their strategies or they can use Global SHAP values per score to make it even more targeted.

3.3 Utilizing Interpretability Techniques for Market Trend Analysis

The adoption of new underwriting models significantly influences market trends, particularly in terms of the market segments that an insurance company leverages. To effectively track these shifts, an automated window-based technique for change detection can be employed. The following steps outline a suggested automated comparative analysis framework:

1. SHAP values of the testing set is saved as a reference dataset along with the predicted risk scores.
2. In the post-production period, proper windows should be selected considering the sample size, seasonality and other patterns. Then, for each window and specific risk scores, an automated t-test should be run between those sets and a respective sample from the reference set [9].

Automating this analysis allows for the systematic identification of model failures or distributional changes arising from the implementation of new underwriting methods. As an illustrative example, if the analysis detects that mean SHAP value for the variable "proportion of claims by region" has shifted to the right significantly (actually having predominantly positive cases, for comparison see reference SHAP values in Fig. 3.), it can indicate a competitive pricing from other insurers and a resultant customer shift. Such insights are crucial for guiding the necessary adjustments during the model retraining process to mitigate undesired market shifts. Future work will expand on this approach, refining and validating the automated methods to foster a robust, continuous analysis of market trends.

4 Conclusion

Drawing upon the literature and after applying different ML and interpretability algorithms to real-life insurance data, the study concludes that Black-Box algorithms generally outperform linear models in accuracy. It highlights the need for tailored interpretability methods to meet the diverse requirements of different stakeholders in the insurance sector. SHAP values are identified as the most effective means to interpret Black-Box models, catering to specific stakeholder needs and useful for monitoring market shifts to inform retraining processes. The study suggests future research should develop a systematic approach to model interpretability, potentially through an interactive dashboard customized for various stakeholders.

References

1. Noordhoek, D.: Regulation of Artificial Intelligence in Insurance: Balancing consumer protection and innovation. The Geneva Association (2023)
2. Diana, A., Griffin, J., Oberoi, J., Yao, J.: Machine-Learning Methods for Insurance Applications. Society of Actuaries, 475 N. Martingale Road, Suite 600, Schaumburg, Illinois 60173 (2019)
3. Lozano-Murcia, C., Romero, F.P., Serrano-Guerrero, J., Peralta, A., Olivas, J.A. Potential Applications of Explainable Artificial Intelligence to Actuarial Problems. Mathematics **12**(5), 635 (2024)
4. Ghonge, M.M., Pradeep, N., Jhanjhi, N.Z., Kulkarni, P.M.: Advances in Explainable AI. Applications for Smart Cities, pp. 1–506 (2024)
5. Chkoniya, V.: Handbook of Research on Applied Data Science and Artificial Intelligence in Business and Industry, pp. 1–626 (2021)
6. Hassija, V., Chamola, V., Mahapatra, A., Singal, A., Goel, D., Huang, K., Scardapane, S., Spinelli, I., Mahmud, M., Hussain, A: Interpreting Black-Box Models: A Review on Explainable Artificial Intelligence. Cogn Comput 16, pp. 45–76 (2024)
7. Eling M., Nuessle D., Staubli, J.: The impact of artificial intelligence along the insurance value chain and on the insurability of risks. Geneva Pap Risk Insur Issues Pract **47**, pp. 205–241 (2022)
8. International Actuarial Association: Actuarial Function (2023)
9. Brodsky B.: Change-Point Analysis in Nonstationary Stochastic Models. Routledge-Chapman & Hall/CRC press (2022)

In October 2024, researchers, students, and practitioners from Armenia, Chile, Germany, and Japan met at the American University of Armenia for the fourth edition of the Cadassca Workshop on Collaborative Technologies and Data Science in Smart City Applications. This book presents their contributions to emerging methodologies in data science, machine learning, applied and human-centered computing.

Logos Verlag Berlin

ISBN 978-3-8325-5855-0