Annals of Computer Science and Information Systems
Volume 36

Position Papers of the 18th Conference on Computer Science and Intelligence Systems

September 17–20, 2023. Warsaw, Poland

Maria Ganzha, Leszek Maciaszek, Marcin Paprzycki,
Dominik Ślęzak (eds.)

PTI

Annals of Computer Science and Information Systems, Volume 36

Series editors:
Maria Ganzha (Editor-in-Chief),
Systems Research Institute Polish Academy of Sciences and Warsaw University of Technology, Poland
Leszek Maciaszek,
Macquarie University, Australia and Wrocław Universty of Economy, Poland
Marcin Paprzycki,
Systems Research Institute Polish Academy of Sciences and Management Academy, Poland
Dominik Ślęzak,
University of Warsaw, Poland and QED Software, Poland, and DeepSeas, USA

Senior Editorial Board:
Wil van der Aalst,
Technische Universiteit Eindhoven (TU/e), Netherlands
Enrique Alba,
University of Málaga, Spain
Marco Aiello,
University of Groningen, Netherlands
Mohammed Atiquzzaman,
University of Oklahoma, USA
Christian Blum,
Artificial Intelligence Research Institute (IIIA-CSIC), Spain
Jan Bosch,
Chalmers University of Technology, Sweden
George Boustras,
European University, Cyprus
Barrett Bryant,
University of North Texas, USA
Rajkumar Buyya,
University of Melbourne, Australia
Chris Cornelis,
Ghent University, Belgium
Hristo Djidjev,
Los Alamos National Laboratory, USA and Institute of Information and Communication Technologies, Bulgaria
Włodzisław Duch,
Nicolaus Copernicus University, Toruń, Poland
Hans-George Fill,
University of Fribourg, Switzerland
Ana Fred,
Technical University of Lisbon, Portugal
Giancarlo Guizzardi,
Free University of Bolzano-Bozen, Italy
Francisco Herrera,
University of Granada, Spain
Mike Hinchey,
University of Limerick, Ireland
Janusz Kacprzyk,
Systems Research Institute, Polish Academy of Sciences, Warsaw, Poland

Irwin King,
The Chinese University of Hong Kong, China
Michael Luck,
King's College London, London, United Kingdom
Ivan Luković,
University of Belgrade, Serbia
Stan Matwin,
Dalhousie University, University of Ottawa, Canada and Institute of Computer Science, Polish Academy of Science, Poland
Marjan Mernik,
University of Maribor, Slovenia
Michael Segal,
Ben-Gurion University of the Negev, Israel
Andrzej Skowron,
the University of Warsaw, Poland
John F. Sowa,
VivoMind Research, LLC, USA
George Spanoudakis,
University of London, United Kingdom

Editorial Associates:
Katarzyna Wasielewska,
Systems Research Institute Polish Academy of Sciences, Poland
Paweł Sitek,
Kielce University of Technology, Poland

TEXnical editor: Aleksander Denisiuk,
University of Warmia and Mazury in Olsztyn, Poland

Promotion and Marketing: Anastasiya Danilenka,
Warsaw University of Technology, Poland

Position Papers of the 18th Conference on Computer Science and Intelligence Systems

Maria Ganzha, Leszek Maciaszek, Marcin Paprzycki, Dominik Ślęzak (eds.)

2023, Warszawa, Polskie Towarzystwo Informatyczne

Annals of Computer Science and Information Systems, Volume 36

Position Papers of the 18th Conference on Computer Science and Intelligence Systems

USB: ISBN 978-83-969601-2-2
WEB: ISBN 978-83-969601-1-5

ISSN 2300-5963
DOI 10.15439/978-83-969601-1-5

© 2023, Polskie Towarzystwo Informatyczne
Ul. Solec 38/103
00-394 Warsaw
Poland

Contact: secretariat@fedcsis.org
http://annals-csis.org/

Cover photo:
Helena Wojciechowska,
Elbląg, Poland

Also in this series:

Volume 37: Communication Papers of the 18th Conference on Computer Science and Intelligence Systems, **ISBN WEB: 978-83-969601-3-9, ISBN USB: 978-83-969601-4-6**
Volume 35: Proceedings of the 18th Conference on Computer Science and Intelligence Systems, **ISBN WEB 978-83-967447-8-4, ISBN USB 978-83-967447-9-1, ISBN ART 978-83-969601-0-8**
Volume 34: Proceedings of the Third International Conference on Research in Management and Technovation **ISBN 978-83-965897-8-1**
Volume 33: Proceedings of the Seventh International Conference on Research in Intelligent and Computing in Engineering, **ISBN WEB: 978-83-965897-6-7, ISBN USB: 978-83-965897-7-4**
Volume 32: Communication Papers of the 17th Conference on Computer Science and Intelligence Systems, **ISBN WEB: 978-83-965897-4-3, ISBN USB: 978-83-965897-5-0**
Volume 31: Position Papers of the 17th Conference on Computer Science and Intelligence Systems, **ISBN WEB: 978-83-965897-2-9, ISBN USB: 978-83-965897-3-6**
Volume 30: Proceedings of the 17th Conference on Computer Science and Intelligence Systems, **ISBN WEB: 978-83-962423-9-6, ISBN USB: 978-83-965897-0-5**
Volume 29: Recent Advances in Business Analytics. Selected papers of the 2021 KNOWCON-NSAIS workshop on Business Analytics**ISBN WEB: 978-83-962423-7-2, ISBN USB: 978-83-962423-6-5**
Volume 28: Proceedings of the 2021 International Conference on Research in Management & Technovation, **ISBN WEB: 978-83-962423-4-1, ISBN USB: 978-83-962423-5-8**
Volume 27: Proceedings of the Sixth International Conference on Research in Intelligent and Computing in Engineering, **ISBN WEB: 978-83-962423-2-7, ISBN USB: 978-83-962423-3-4**
Volume 26: Position and Communication Papers of the 16th Conference on Computer Science and Intelligence Systems, **ISBN WEB: 978-83-959183-9-1, ISBN USB: 978-83-962423-0-3**
Volume 25: Proceedings of the 16th Conference on Computer Science and Intelligence Systems, **ISBN WEB 978-83-959183-6-0, ISBN USB 978-83-959183-7-7, ISBN ART 978-83-959183-8-4**

DEAR Reader, it is our pleasure to present to you Position Papers of the 18th Conference on Computer Science and Intelligence Systems (FedCSIS 2023), which took place in Warsaw, Poland, on September 17-20, 2023.

Position papers comprise two categories of contributions – challenge papers and emerging research papers. *Challenge papers* propose and describe research challenges in theory, or practice, of computer science and intelligence systems. Papers in this category are based on deep understanding of existing research or industrial problems. Based on such understanding and experience, they define new exciting research directions and show why these directions are crucial to the society at large. *Emerging research papers* present preliminary research results from work-in-progress, based on sound scientific approach but presenting work not completely validated as yet. They describe precisely the research problem and its rationale. They also define the intended future work including the expected benefits from solution to the tackled problem. Subsequently, they may be more conceptual than experimental.

FedCSIS 2023 was chaired by Jarosław Arabas and Sławomir Zadrożny, while Przemysław Biecek acted as the Chair of the Organizing Committee. This year, FedCSIS was organized by Polish Information Processing Society (Mazovia Chapter), IEEE Poland Section Computer Society Chapter, Systems Research Institute of Polish Academy of Sciences, as well as Faculty of Electronics and Information Technology and Faculty of Mathematics and Information Sciences, of Warsaw University of Technology.

FedCSIS 2023 was technically co-sponsored by IEEE Poland Section, IEEE Czechoslovakia Section Computer Society Chapter, IEEE Poland Section Systems, Man, and Cybernetics Society Chapter, IEEE Poland Section Computational Intelligence Society Chapter, Committee of Computer Science of Polish Academy of Sciences, and Mazovia Cluster ICT. Moreover, two years ago, the FedCSIS conference series formed strategic alliance with QED Software, a Polish software company developing AI-based products, and this collaboration has been continued.

In 2023, FedCSIS was sponsored by QED Software, Samsung, Hewlett Packard Enterprise, Łukasiewicz Research Network – Institute of Innovative Technologies EMAG, MDPI, Sages, Efigo, and CloudFerro.

During FedCSIS 2023, the keynote lectures were delivered by:

- Lipika Dey, Tata Consultancy Services, India, keynote title: *Deciphering Clinical Narratives – Augmented Intelligence for Decision Making in Health Care Sector*
- Marta Kwiatkowska, University of Oxford, United Kingdom, keynote title: *When to trust AI…*
- Andrea Omicini, Alma Mater Studiorum – Università di Bologna, Italy, keynote title: *Measuring Trustworthiness in Neuro-Symbolic Integration*
- Roman Słowiński, Poznań University of Technology, Poland, keynote title: *Multiple Criteria Decision Aiding by Constructive Preference Learning*

Moreover, two special guests delivered invited presentations:

- Gianpiero Cattaneo, Retired from Department of Informatics, Systems and Communications, University of Milano-Bicocca, Italy, invited presentation title: *Abstract Approach to Entropy and Co-Entropy in Measurable and Probability Spaces*
- Jerzy Nawrocki, Poznań University of Technology, Poland, invited presentation title: *Towards reliable rule mining about code smells: The McPython approach*

FedCSIS 2023 consisted of Main Track, with five Topical Areas and Thematic Tracks. Some of Thematic Tracks have been associated with the FedCSIS conference series for many years, while some of them are relatively new. The role of Thematic Tracks is to focus and enrich discussions on selected areas, pertinent to the general scope of the conference.

Each contribution, found in this volume, was refereed by at least two referees. They are presented in alphabetic order, according to the last name of the first author. The specific Topical Area or Thematic Track that given contribution was associated with is listed in the article metadata.

Making FedCSIS 2023 happen required a dedicated effort of many people. We would like to express our warmest gratitude to the members of Senior Program Committee, Topical Area Curators, Thematic Track Organizers and to members of FedCSIS 2023 Program Committee. In particular, we would like to thank those colleagues who have refereed all of the 358 submissions.

We thank the authors of papers for their great contributions to the theory and practice of computer science and intelligence systems. We are grateful to the keynote and invited speakers for sharing their knowledge and wisdom with the participants.

Last, but not least, we thank Jarosław Arabas, Sławomir Zadrożny, and Przemysław Biecek. We are very grateful for all your efforts!

We hope that you had an inspiring conference. We also hope to meet you again for the 19th Conference on Computer Science and Intelligence Systems (FedCSIS 2024), which will take place in Belgrade, Serbia, on September 8-11, 2024.

Co-Chairs of the FedCSIS Conference Series:

Maria Ganzha, *Warsaw University of Technology and Systems Research Institute Polish Academy of Sciences, Poland*

Leszek Maciaszek (Honorary Chair), *Macquarie University, Australia and Wrocław University of Economics, Poland*

Marcin Paprzycki, *Systems Research Institute Polish Academy of Sciences, and Warsaw University of Management, Poland*

Dominik Ślęzak, *University of Warsaw, Poland and QED Software, Poland, and DeepSeas, USA*

Annals of Computer Science and Information Systems, Volume 36

Position Papers of the 18th Conference on Computer Science and Intelligence Systems

September 17–20, 2023. Warsaw, Poland

TABLE OF CONTENTS

POSITION PAPERS

Position Papers of the 18th Conference on Computer Science and Intelligence Systems pp. 1–8
DOI: 10.15439/2023F7902
ISSN 2300-5963 ACSIS, Vol. 36

Large Minded Reasoners for Soft and Hard Cluster Validation – Some Directions

1st MANI A
Machine Intelligence Unit
Indian Statistical Institute, Kolkata, India
Email: amani.rough@isical.ac.in, a.mani.cms@gmail.com
ORCID: 0000-0002-0880-1035

2nd SUSHMITA MITRA
Machine Intelligence Unit
Indian Statistical Institute, Kolkata, India
Email: sushmita@isical.ac.in
ORCID: 0000-0001-9285-1117

***Abstract*—In recent research, validation methods for soft and hard clustering through general granular rough clusters are proposed by the first author. Large-minded reasoners are introduced and studied in the context of new concepts of non-stochastic rough randomness in a separate paper by her. In this research, the methodologies are reviewed and new low-cost scalable methodologies and algorithms are invented for computing granular rough approximations of soft clusters for many classes of partially ordered datasets. Specifically, these are applicable to datasets in which attribute values are numeric, vector valued, lattice-ordered or partially ordered. Additionally, new research directions are indicated.**

***Index Terms*—Soft Clustering, General Rough Sets, Cluster Validation, Rough Randomness, Non-Stochastic Randomness, Axiomatic Granules, Large Minded Reasoners, Ontology of Clustering, Low Cost Computing, Tolerances, Clean Rough Randomness**

I. Introduction

BOTH hard and soft clustering with or without additional semi-supervision are popular tools for classification in the AIML literature [1]. The problems of cluster validation is a long-standing one. Numeric measures are known to be unreliable, inconsistent and mathematically unjustifiable [2], [3]. Known proofs often proceed from questionable statistical and topological assumptions [1], [4] about the context associated with a dataset. Recent attempts to solve the problem from an axiomatic granular rough set perspective [5], [6] are proposed by the first author in the paper [7]. Ortho-partitions are related to three-way clusterings in other recent work [8]. In the present paper, the focus is on clustering contexts that can be coerced into granular tolerance frameworks. It is not necessary that the clustering be distance-based.

Suppose a soft clustering $\mathcal{S}$ is defined as a finite sequence of ordered pairs of cores and their fringes. The underlying philosophy of the invented methodology is that $\mathcal{S}$ is valid relative to a granular rough model $\mathcal{R}$ if and only if the components of $\mathcal{S}$ are definite objects or are very closely approximated in $\mathcal{R}$.

In the earlier work mentioned [7], a rough model is essential for cluster validation. The meaning associated with the construction of the approximations is the basis of the framework. For example, blocks of a tolerance [9]–[11] can be interpreted as maximal sets of mutually similar objects, and approximations formed as unions of blocks have a simple disjunctive meaning (over higher order concepts of similarity). However, if no rough models seem to be reasonable, then can they be discovered/constructed? A solution for this problem is proposed through a slightly lengthy process involving concepts of clean rough randomness, large minded reasoners, and invented algorithms based on recent advances in the theory of tolerances. Research directions are additionally provided.

The first author's work is supported by grant no. SR/WOS-A/PM-22/2019 of DST, India

A. Structure of this Paper

Some background is provided in the next section. The concept of rough randomness is explained from a different perspective, the associated polysemy is fixed, and large minded reasoners that embody clean rough randomness are formalized in the third section. The overall strategy for validation of soft and hard clustering is formulated next. In the fifth, algorithms for validation methods for truly unsupervised clustering are invented. Related directions and applications are discussed in the last section.

II. Background

A *distance function* on a set S is a function $\rho : S^2 \longmapsto \Re_+$ that satisfies

$$(\forall a)\rho(a,a) = 0 \qquad \text{(distance)}$$

The collection $\mathcal{B} = \{B_\rho(x,r) : x \in S \& r > 0\}$ of all r-spheres generated by ρ is a weak base for the topology τ_ρ defined by

$$V \in \tau_\rho \text{ if and only if } (\forall x \in V \exists r > 0) B_\rho(x,r) \subseteq V$$

Any $\epsilon > 0$ and a distance function ρ determines a tolerance T defined by

$$Tab \text{ if and only if } \rho(a,b) + \rho(b,a) \leq \epsilon.$$

One can define other tolerances through conditions such as

$$\frac{\rho(a,b) + \rho(b,a)}{1 + \rho(a,b) + \rho(b,a)} \leq \epsilon.$$

The point is that a function much weaker than a semimetric suffices for defining a tolerance relation. More complex definitions are often possible.

©2023, PTI

Thematic track: Rough Sets: Theory and Applications

Proposition 1. *For a numeric complete information table $\mathcal{I}$, the following holds:*

1) *Valuations for each attribute are totally ordered by $\leq$,*
2) *$\mathcal{O}$ is totally ordered relative to the induced lexicographic order.*
3) *$\mathcal{O}$ is lattice ordered relative to $\leq$ defined by $(a_1, \ldots, a_n) \preceq (b_1, \ldots, b_n)$ if and only if $\&_i a_i \leq b_i$ with $a_i, b_i \in Ran(\nu(, At_i)$.*

However, a numeric table is not necessary for any of the three properties to hold.

A. Tolerances

For more details, the reader is referred to the works [9], [12], [13].

If T is a tolerance on a set S, then a *pre-block* of T is a subset $K \subseteq S$ that satisfies $K^2 \subseteq T$. The set of all pre-blocks of T is denoted by $p\mathcal{B}(T)$. Maximal pre-blocks of T with respect to the inclusion order are referred to as *blocks*. The set of all blocks of T is denoted by $\mathcal{B}(T)$. If $S = \langle underline S,\, f_1,\, f_2,\, \ldots,\, f_n, (r_1,\, \ldots,\, r_n)\rangle$ ($\underline{S}$ being a set and f_i being r_i-place operation symbols interpreted on it) is an algebra, then a tolerance T is said to be *compatible* if and only if for each $i \in \{1, 2, \ldots n < \infty\}$,

$$\&_{j=1}^{r_i} T a_j b_j \longrightarrow T f_i(a_1, a_2, \ldots a_{r_i}) f_i(b_1, b_2, \ldots b_{r_i}).$$

When S is a lattice, every tolerance is the image of a congruence by a surjective morphism : $S \longmapsto S$. Further, if $A, B \in \mathcal{B}(T)$, then $\{a \vee b : a \in A \& b \in B\}, \{a \wedge b : a \in A \& b \in B\} \in p\mathcal{B}(T)$. The smallest blocks containing these are unique, and the resulting lattice of blocks is denoted by $S|T$. The set $\mathbf{UBD}(S) = \{\mathcal{B}(T) : T \in Tol(S)\}$ will be referred to as the *universal block distribution* (UBD) of S. It can be assigned the same algebraic lattice order on $Tol(S)$.

A sublattice Z of a lattice S is called a *convex sublattice* if and only if it satisfies $(\forall x, b \in Z)(x \leq a \leq b \longrightarrow a \in Z)$. The blocks of a lattice are all convex sublattices. If C is a subset of S then $\downarrow C$, and $\uparrow C$ will respectively denote the lattice-ideal and lattice-filter generated by C. The following result [12], [14], [15] is not usable for a direct computational strategy:

Theorem 1. *For a finite lattice L, a collection $\mathcal{C}$ of nonempty subsets is the set of all blocks of a tolerance $T \in Tol(L)$ if and only if it is a collection of intervals of the form $\{[a_i, b_i] : i \in I\}$, and*

- $\bigcup_{i \in I} [a_i, b_i] = L$
- *For all $i, j \in I$, $(a_i = a_j \longrightarrow b_i = b_j)$.*
- $(\forall i, j \in I)(\exists k \in I)\, a_k = a_i \vee a_j \,\&\, b_i \vee b_j \leq b_k$.

Theorem 2. *In the context of Theorem 1,*

1) $(\forall C, E \in \mathcal{C})\, (\downarrow C = \downarrow E \Longleftrightarrow \uparrow C = \uparrow E)$.
2) *For any two elements $C, A \in \mathcal{C}$ there exist E, F such that $(\downarrow C \vee \downarrow A) = \downarrow E$, $(\uparrow C \vee \uparrow A) \leq \uparrow E$, $\downarrow F \leq (\downarrow A \wedge \downarrow C)$, and $(\uparrow C \wedge \uparrow A) = \uparrow F$).*

For finite chains, the following holds [16]

Theorem 3. 1) *A collection $\mathcal{C}$ of subsets of the chain $L_n = \langle \{0, 1, 2, \ldots n-1\}, \leq \rangle$ is the set of all blocks of a tolerance $T \in Tol(L)$ if and only if $\mathcal{C}$ is of the form $\{[n_i, m_i] : i = 1, \ldots k\}$ for some $1 \leq k \leq n-1$, with $n_1 = 0$, $m_k = n-1$, and $n_i < n_{i+1} \leq m_i + 1$, and $m_i < m_{i+1}$ for all $i = 1, \ldots k$.*

2) *A collection $\mathcal{C}$ of subsets of the chain $L_n = \langle \{0, 1, 2, \ldots n-1\}, \leq \rangle$ is the set of all blocks of a glued tolerance $T \in Glu(L)$ if and only if $\mathcal{C}$ is of the form $\{[n_i, m_i] : i = 1, \ldots k\}$ for some $1 \leq k \leq n-1$, with $n_1 = 0$, $m_k = n-1$, and $n_i < n_{i+1} \leq m_i < m_i + 1$, and $m_i < m_{i+1}$ for all $i = 1, \ldots k$.*

3) *A collection $\mathcal{C}$ of subsets of the chain $L_n = \langle \{0, 1, 2, \ldots n-1\}, \leq \rangle$ is the set of all blocks of a congruence $R \in Con(L)$ if and only if $\mathcal{C}$ is of the form $\{[n_i, m_i] : i = 1, \ldots k\}$ for some $1 \leq k \leq n-1$, with $n_1 = 0$, $m_k = n-1$, and $n_i < n_{i+1} = m_i + 1$, and $m_i < m_{i+1}$ for all $i = 1, \ldots k$.*

The next theorem is a combination of special cases of known results.

Theorem 4.

- *Tolerances on a product of finite lattices are directly decomposable [17].*
- *If a lattice L is a direct product $\Pi_{i=1}^n L_i$ of the lattices L_i, then $Tol(L) \simeq Tol(L_1) \times Tol(L_2) \times \ldots Tol(L_n)$*

So every $T \in Tol(L)$ can be written as a direct product of tolerances $T_i \in Tol(L_i)$ $(i = 1, 2, \ldots, n)$. That is

$$T = \Pi_{i=1}^n T_i = \{(a, b);\, (e_i a, e_i b) \in T_i, \text{ for } i = 1, 2 \ldots n\}.$$

Further,

Theorem 5. *Let S_1 and S_2 be two lattices with compatible tolerances T_1 and T_2 respectively. If $T(a_1, a_2, \ldots a_n)(b_1, b_2, \ldots b_n)$ if and only if $T_1 a_1 b_1 \& T_2 a_2 b_2 \& \ldots T_n a_n b_n$, then the blocks of T are direct products of the blocks of the component tolerances.*

Proof. Suppose $\mathcal{S}_\infty = \{B_i i \in [1, n]\}$ and $\mathcal{S}_\in = \{F_i i \in [1, m]\}$ are two distinct normal covers for the same tolerance T on an algebra S. If Tab for any $a, b \in S$, then it is necessary that $a, b \in F_i$, and $a, b \in B_j$ for some i, j. This means $\mathcal{S}_\infty$ must be a mere rearrangement of $\mathcal{S}_\in$. In other words, normal covers of a tolerance are unique. □

The above proof works for finite direct products as well. It means that one needs to create all permutations of the blocks on components in general.

B. Approximations

Let $S = \langle \underline{S}, T \rangle$ be a general approximation space, with $\underline{S}$ being a set, and T a tolerance relation on it. If $\mathcal{G}$ is the set of all blocks of T, and $A \subseteq S$, then the following granular approximations [9] will be used in the main algorithms (the semantics and history of the approximations are described in the mentioned reference).

$$A^l = \bigcup\{H : H \subseteq A,\, H \in \mathcal{G}\} \quad \text{(Lower)}$$

$$A^u = \bigcup\{H : A \cap H \neq \emptyset,\, H \in \mathcal{G}\} \quad \text{(Upper)}$$

$$A^{u_b} = A^u \setminus A^{cl} \quad \text{(Bited Upper)}$$

Further, the model $\mathbb{S} = \left\langle \underline{\wp(\mathbb{S})}, \gamma, l, u, \subseteq, \cup, \cap, \emptyset, S \right\rangle$ generated by the granular approximations on the power set $\underline{\wp(\mathbb{S})}$ will be used to discuss cluster validation. It is additionally a set HGOS in the sense of the first author [6].

C. Validation Indices – a Brief Critique

The process or concept of *cluster validation* generally refers to exploring the quality of one or more clustering methods and possibly comparing them. In almost all cases, true class information is not available (that is if one avoids looking at anything apart from the dataset) and validation methods are inherently not rigorous even in comparison to statistical methods used in supervised learning. For example, in a regression modeling context, it is possible to say something concrete about model fit relative to a set of statistical assumptions (that may be invalid). Clustering contexts are difficult to investigate from similar perspective. In this subsection, some issues faced are mentioned.

A number of indices for quantifying a clustering's quality are known. Typically, they are used to simply assess the quality of a single clustering or to select the most appropriate clustering method and related parameters (like number of optimal clusters). For datasets of the form $\{x_i\}_{i=1}^n$ over a Euclidean space with standard norm or some distance ρ, common indices such as the Davies-Bouldin (DB) index, Calinski–Harabasz(CH) index and variants thereof start with measures of within-cluster variation and compare them with measures of between-cluster variation from a numerical perspective over the real number field. Many adaptations to rough clustering are also known [18], [19]. In the case of CH-index, the shape of associated clusters are assumed to be spherical, with data points concentrated around the cluster means. Even if points close to the boundary of the sphere are close to points on the boundary of another cluster, the distance between the two clusters will be the distance between the means. Further, the index is naturally connected with the hard k-means algorithm. These can be used in determining the appropriateness of the index in a specific application context. From this *it should be clear that indices carry very little information about ontology*.

Latent class model-based approach (LCC) is sometimes used for clustering multivariate categorical data. There each cluster is assumed to be a mixture component and the whole is a mixture of probability distributions [20]. These are not well-related to distance based approaches, though the *average silhouette width* (ASW) measure that emphasizes the separation between clusters and their neighbors is known to be useful. ASW is also heavily used in fuzzy clustering.

External Criteria: A simple example of an external criterion for hard clustering is the *quality index* Q_+. If N_+, N_- and N are respectively the number of correctly assigned, incorrectly assigned and total number of clusters.

$$Q_+ = \frac{N_+}{N} \,\&\, Q_- = \frac{N_-}{N}.$$

This measure can be generalized to rough clustering [19], [21], and other soft approaches (with cores and fringes). Obviously these indices avoid most of the complexity and semantics involved in the clustering process. However, they allow gradation of boundaries, and are less controversial than other indices because of the minimal number of assumptions.

In general, the following remarks about indices may be noted:

- Cluster validation is sometimes done from a statistical test perspective. The null hypothesis is taken to be the statement that the data homogeneous and unclustered according to a null model. Reasonable clusterings are expected to be significantly better than what is expected relative to its performance on the null model. This is usually done relative to specific cluster performance indices. Both the reality of the statistical scheme of things assumed and indices used remain very questionable.
- Clusters with complicated shapes are common in application contexts like image processing. For example, in many photos of natural scenarios, similar leaves can be in different parts [22] (these are handled with descriptive proximities and related functions). Indices for clusters in such contexts are not well-developed.
- Combining multiple validation indices for the purpose of measuring multiple characteristics has limited scope and the act of combining does not go beyond forming a set of indices [3].

In hierarchical clustering in particular, indices such as partial R-squared monotonically change with number of clusters. Strong decrease (or increase) followed by weak decrease (or increase) of the index relative to the number of clusters correspond to their optimal values. Further, the applicability of associated indices is limited.

It can therefore be asserted that *most indices assume some heuristics that are not well understood and in some cases even the values produced may not be clear (see [2], [3]) for details*. From a mathematical point of view, a few rigorous studies on indices in the context of semimetric based clustering are known.

III. Clean Rough Randomness and LMR

Many types of randomness are known in the literature. Stochastic randomness, often referred to as randomness, is often misused without proper justification. In the paper [23], a phenomenon is defined to be *stochastically random* if it has probabilistic regularity in the absence of other types of regularity. In this definition, the concept of regularity may be understood as *mathematical regularity* in some sense. Generalizations of mathematical probability theory through hybridization with rough sets from a stochastic perspective are explained in the book [24]. This approach is not ontologically

consistent with pure rough reasoning or explainable AI as its focus is on modeling the result of numeric simplifications in a measure-theoretic context.

A *rare property* in the theory of computation is an effectively testable property that is valid over a set of measure zero. A finite or infinite sequence $\mathbf{x}$ is said to be *algorithmically random* if and only if no computational agent recognizes $\mathbf{x}$ as possessing some rare property (for details see [23], [25]). While associations with subjective probability are known, connections of such ideas with rough sets are not known in the literature.

Empirical studies show that humans cannot estimate measures of stochastic randomness and weakenings thereof in real life properly [26]. This is consistent with the observation that connections in the rough set literature between specific versions of rough sets and subjective probability theories (Bayesian or frequentist) are not good approximations. In fact, rough inferences are grounded in some non-stochastic comprehension of attributes (their relation with the approximated object in terms of number or relative quantity and quality) [27], [28].

The idea of *rough randomness* is expressed by the first author [29] as follows: *a phenomenon is roughly random if it can be modeled by general rough sets or a derived process thereof*. To avoid the resulting polysemy (as the term is used in a different sense in the monograph [24]), it is useful to rename it as *clean rough randomness* (or C-rough randomness for short). In concrete situations, such a concept should be realizable in terms of C-roughly random functions or predicates defined below (more variations will be part of future work):

Definition 1. *Let $\mathcal{A}_\tau$ be a collection of approximations of type τ, and E a collection of rough objects [9] defined on the same universe S, then by a* C-rough random function of type-1 *(CRRF1) will be meant a partial function*

$$\xi : \mathcal{A}_\tau \longmapsto E.$$

Definition 2. *Let $\mathcal{A}_\tau$ be a collection of approximations of type τ, $\mathcal{S}$ a subset of $\wp(S)$, and $\Re$ the set of reals, then by a* C-rough random function of type-2 *(CRRF2) will be meant a function*

$$\chi : \mathcal{A}_\tau \times \mathcal{S} \longmapsto \Re.$$

Definition 3. *Let $\mathcal{A}_\tau$ be a collection of approximations of type τ, and F a collection of objects defined on the same universe S, then by a* C-rough random function of type-3 *(CRRF3) will be meant a function*

$$\mu : \mathcal{A}_\tau \longmapsto F.$$

Definition 4. *Let $\mathcal{O}_\tau$ be a collection of approximation operators of type τ_l or τ_u, and E a collection of rough objects defined on the same universe S, then by a* C-rough random function of type-H *(CRRFH) will be meant a partial function*

$$\xi : \mathcal{O}_\tau \times \wp(S) \longmapsto E.$$

It is obvious that a CRRF1 and CRRF2 are independent concepts, while a total CRRF1 is an CRRF3, and CRRFH is distinct (though related to CRRF3). The set of all such functions will respectively be denoted by $CRRF1(S, E, \tau)$, $CRRF2(S, \Re, \tau)$, $CRRF3(S, F, \tau)$, and $CRRFH(S, E, \tau)$. Examples that show the semantic nature of the associations are mentioned below:

Examples: CRRF

Example 1. *Let S be a set with a pair of lower (l) and upper (u) approximations satisfying (for any $a, b, x \subseteq S$)*

$$x^l \subseteq x^u \quad \text{(int-cl)}$$
$$x^{ll} \subseteq x^l \quad \text{(l-id)}$$
$$a \subseteq b \longrightarrow a^l \subseteq b^l \quad \text{(l-mo)}$$
$$a \subseteq b \longrightarrow a^u \subseteq b^u \quad \text{(u-mo)}$$
$$\emptyset^l = \emptyset \quad \text{(l-bot)}$$
$$S^u = S \quad \text{(u-top)}$$

The above axioms are minimalist, and most general approaches satisfy them.

In addition, let

$$\mathcal{A}_\tau = \{x : (\exists a \subseteq S)\, x = a^l \text{ or } x = a^u \quad (1)$$
$$E_1 = \{(a^l, a^u) : a \in S\} \quad \text{(E1)}$$
$$F = \{a : a \subseteq S \& \neg\exists b b^l = a \vee b^u = a\} \quad \text{(E0)}$$
$$E_2 = \{b : b^u = b \& b \subseteq S\} \quad \text{(E2)}$$
$$\xi_1(a) = (a, b^u) \text{ for some } b \subseteq S \quad \text{(xi1)}$$
$$\xi_2(a) = (b^l, a) \text{ for some } b \subseteq S \quad \text{(xi2)}$$
$$\xi_3(a) = (e, f) \in E_1 \,\&\, e = a \text{ or } f = a \quad \text{(xi3)}$$

E_1 in the above is a set of rough objects, and a number of algebraic models are associated with it [9]. A partial function $f : \mathcal{A}_\tau \longmapsto E_1$ that associates $a \in \mathcal{A}_\tau$ with a minimal element of E_1 that covers it in the inclusion order is a CRRF of type 1. For general rough sets, this CRRF can be used to define algebraic models and explore duality issues [13], and for many cases associated these are not investigated. A number of similar maps with value in understanding models [27] can be defined. Rough objects are defined and interpreted in a number of other ways including F or E_2.

Conditions xi1-xi3 may additionally involve constraints on b, e and f. For example, it can be required that there is no other lower or upper approximation included between the pair or that the second component is a minimal approximation covering the first. It is easy to see that

Theorem 6. *ξ_i for $i = 1, 2, 3$ are CRRF of type-1.*

Example 2. *In the context of the above example, rough inclusion functions, membership, and quality of approximation functions [30], [31] can be used to define CRRF2s. An example is the function ξ_5 defined by*

$$\xi_5(a, b) = \frac{Card(b \setminus a)}{Card(b)} \quad \text{(III.1)}$$

In the paper [29], it is additionally proved that

Theorem 7. *A rough random variable [24] in the sense of Liu, is not a rough random function of any type.*

IV. Hard and Soft Clustering Validation Strategies

The considerations of this section will be restricted to validation of soft clustering defined through ortho-pairs [8]. In the mentioned paper, the authors do not explicitly say that their universe is finite, and it is not clarified whether it is a semimetric set (a set with a semimetric) or a semimetric space. The former does not always define a semimetric topology. Connections with proximities [32] are additionally not mentioned. *However, these assumptions are not required for obtaining three-way clusters or rough clusters in their sense*.

An ortho pair is a pair of disjoint subsets (of a universe S) of the form $O = (C, F)$ with C being the core and F being the boundary or fringe that satisfies $C \cap F = \emptyset$. An ortho-partition $\mathcal{O}$ is a collection of ortho pairs of the form

$$\{(C_1, F_1),\, (C_2, F_2),\, (C_3, F_3),\, \ldots, (C_n, F_n)\}$$

that satisfies O0, O1, O2, and O3

$$\text{For all } i\; C_i \neq \emptyset \qquad \text{(T1)}$$

$$\text{For all } i\; C_i \cap F_i = \emptyset \qquad \text{(O0)}$$

$$\text{If } i \neq j \text{ then } C_i \cap C_j = C_i \cap F_j = C_j \cap F_i = \emptyset \qquad \text{(O1)}$$

$$\bigcup (C_i \cup F_i) = S \qquad \text{(O2)}$$

$$(\forall x)(x \in F_i \longrightarrow (\exists j) j \neq i \,\&\, x \in F_j) \qquad \text{(O3)}$$

$$\text{If } \forall i\; x \notin C_i \text{ then } \exists i, j\; i \neq j \,\&\, x \in F_i \cap F_j \qquad \text{(R2)}$$

A *rough clustering* is a collection of ortho pairs that satisfies O0, O1, and R2. While a *three-way clustering* is a collection of orthopairs that satisfies O1, O2, and T1. However, it is interpreted as a soft clustering $\mathcal{K}$ in which each cluster K_i is associated with three regions C_i, F_i and $E_i = (C_i \cup F_i)^c$. The last region being interpreted as the *certainly not that* region.

While rough clusterings can be interpreted as ortho-partitions, three-way clusterings are collections of ortho-pairs that do not satisfy O3 in general. However, it is possible to collect the elements not satisfying O3 and create a new cluster – the resulting clustering satisfies O3. Therefore, ortho-partitions suffice for representing semi metric based rough, and three-way clustering, and have a few arguably nice properties (of scale invariance, generalized richness and consistency). If $\mathfrak{D}(S)$ is the set of all semimetrics on S, and $\Pi(S)$ the set of all partitions of S, and $\mathcal{O}(S)$ the set of all ortho partitions on S, then an algorithm is a computable function $c_{tw} : \mathfrak{D}(S) \longmapsto \mathcal{O}(S)$.

Here we are concerned with validation techniques for the clustering. Our basic principle for validation that *if the interpretation of the cores and exteriors are almost the same as their respective approximations in a granular rough semantics, then they are valid relative to the semantics*. This is useful because granular rough semantics in the sense of the first author [6], [9] can explain the meaning of the clusters. Formally,

Definition 5. *Let* $\mathbb{S} = \left\langle \underline{\wp(\mathbb{S})}, \gamma, l, u, \subseteq, \leq, \cup, \cap, \emptyset, S \right\rangle$ *be the set HGOS generated by a tolerance and its granular approximations. Further, let* $\mathcal{Z} = \{(C_i, F_i) i = 1, \ldots r\}$ *be a soft clustering on* S.

- *The* lower deficit *of a soft cluster* $(C, F) \in \mathcal{Z}$ *will be the pair* $((C \setminus C^l)^u, (F \setminus F^l)^u)$,
- *While its* upper deficit *will be the pair* $((C^u \setminus C)^u, (F^u \setminus F)^u)$

The lower and upper deficit of (C, F) *will respectively be denoted by* $(C^\flat, F^\flat))$ *and* $(C^\eth, F^\eth)$. *For hard clustering, it suffices to restrict attention to the core alone.*

Definition 6. *In the context of Definition 5, a soft cluster* $(C, F) \in \mathcal{Z}$ *will be said to*

- lu-valid *if and only if* $C^l = C^u = C$ *and* $F^l = F^u = F$
- l-pre-valid *if and only if* $(\exists V, W \in \mathbb{S}) V^l = C \& W^l = F$.
- u-pre-valid *if and only if* $(\exists V, W \in \mathbb{S}) V^u = C \& W^u = F$.
- l-traceable *if and only if* $(\exists V, W \in \mathbb{S}) V = C^l \& W = F^l$.
- u-traceable *if and only if* $(\exists V, W \in \mathbb{S}) V = C^u \& W = F^u$.

In addition, if all soft clusters in $\mathcal{Z}$ *are* l*-pre-valid (resp.* lu*-valid,* u*-pre-valid,* l*-traceable,* u*-traceable) then* $\mathcal{Z}$ *will itself be said to be* l-pre-valid *(resp.* lu*-valid,* u*-pre-valid,* l*-traceable,* u*-traceable).*

Proposition 2. *In the context of Definition 5, if the* l*-deficit (resp.* u*-deficit) of a hard cluster* C *is computable, then it must necessarily be* l*-traceable (resp.* u*-traceable).*

Proof. If a cluster C has l-deficit A, then it is necessary that $A = (C \setminus C^l)^u$. However, for this C^l should be an element of $\mathbb{S}$. The proof for the u-deficit is similar. □ □

Proposition 3. *In the context of Definition 5, if the* l*-deficit (resp.* u*-deficit) of a soft cluster* (C, F) *is computable, then it must necessarily be* l*-traceable (resp.* u*-traceable).*

The central idea of lu-validity (and weakenings thereof) is that of representability in terms of granules and approximations. These do not test the key predicate δ for validation, and the aspect is left to the process of construction of rough approximations. By contrast, the $*$-deficits are an internal measure of what is lacking or what is in excess.

In relation to the framework of minimal soft clustering system (MSS) invented in the paper [7], it is possible to define the ternary predicate δ through the above concepts. Intended meanings of δabc are a *is closer to* b *than* c in some sense, a *is more similar to* b *than* c *in some sense* and variants thereof. This predicate covers the intent of using metrics, similarities, dissimilarities, proximities, descriptive proximities, kernels and other functions for the purpose.

V. Algorithms and LMR Algorithms

In this section, improved algorithms for the computation of blocks that avoid weakenings are invented. These improve earlier work of the first author [29].

A. Direct Algorithms-1,2

The following two algorithms are used in a forthcoming paper on satellite remote sensing by the first author. They are resource intensive, as the computational ease is limited by the max-clique algorithms. Their complexities are directly defined by that of the similarity matrix computation and the maximal clique algorithms. In the paper, a low-cost implementation could be used through supplementary measures.

Suppose a hard clustering $\{C_i\}_{i=1}^k$ or a soft clustering $\{(C_i, E_i)\}_{i=1}^k$ [1], [8] obtained through any method is given.

Algorithm-1:

Distance: Specify distinct distance function σ_i on the i th column (attribute) for each i

Tolerance: Define a similarity (tolerance relation) T_i on the ith column.

Conjunction: Combine to a single tolerance relation over objects on the table through conjunction of instances across columns.

Relation: Compute the similarity matrix through parallelized methods.

Granules: Compute the blocks of the tolerance by a maximal clique algorithm (for example the modified Bronkerbosch algorithm [33]).

Approximations: Compute granular rough approximations of C_i, and E_i for each i and estimate the closeness of the cluster core or exterior to decide on validation.

If the rough model can explain the soft/hard clustering, then the latter is meaningful and valid.

Algorithm-2:

Distance: Specify a single distance function σ between objects

Tolerance: Define a similarity (tolerance relation) T on the basis of descriptive statistics relative to σ.

Relation: Compute the similarity matrix through parallelized methods.

Granules: Compute the blocks of the tolerance by a maximal clique algorithm.

Approximations: Compute granular rough approximations of C_i, and E_i for each i and estimate the closeness of the cluster core or exterior to decide on validation.

Theorem 8. *The direct algorithm-2 has a computational complexity of* $\mathsf{O}(dk^{3^{d/3}} + N^3)$*, where* d *is the degeneracy of the* n *-vertex graph corresponding to the similarity relation,* k *the number of rows, and* N *is the maximum of number of rows and columns in the dataset.*

Proof. $\mathsf{O}(dk^{3^{d/3}})$ is the complexity of computing max-cliques, while $\mathsf{O}(N^3)$ is that of computing the distance matrix. List operations are assumed to be of linear complexity. □

B. Improved AGRSSA (IAGRSSA)

In an earlier preprint of the first author [29], the AGRSSA (Axiomatic Granular Reversed Similarity Based Semi-Supervised) algorithm(s) was proposed. This is improved below through relaxed assumptions, and stricter constraints on the decision steps involved. It is assumed that each column (attribute) is totally ordered, and that an order-compatible distance (as opposed to a metric) is defined on them. Specifically, it applies to all numeric (real valued) datasets.

IAGRSSA:

Distance: Specify distinct distance functions on each column (attribute).

Quantiles: Identify f-quantiles at a suitable level of precision on each column. Let these be $\{q_{i1}, q_{i2}, \ldots q_{if}\}$ on the ith column based on the distance specified earlier.

Interval Boundaries: Interval boundaries are specifiable by the sequence $\perp_i, q_{i1} - e_{i1}, q_{i1} + e_{i1}, q_{i2} - e_{i2}, q_{i2} + e_{i2}, \ldots, q_{if} - e_{if}, q_{if} + e_{if}, \top_i$. The quantities $e_{i1}, e_{i2}, \ldots e_{if}$ need to computed as a fraction of the measures of variation or other heuristics.

Decision on Blocks: Assume that the intervals on each column are exactly the set of blocks.

Blocks: Form all possible products of the sequence of blocks on each column to form the set of admissible blocks. That is if $\{B_{ij}\}_{j=1}^f$ is the set of blocks on the ith column, the blocks of the whole dataset would have the form $B_{1j_1} \times B_{2j_2} \times \ldots \times B_{nj_k}$, with k being the number of columns and j_i taking values from $1, 2, \ldots f$.

Approximations: Compute granular rough approximations by Subsection II-B and perform decision-making. If a set of objects H are to be approximated, then

1) The lower approximation of H is the union of blocks included in it.
2) The lower approximation of H is the union of blocks that have some common elements with H.

Meaning: This can be specified directly from blocks, or through its associated tolerance.

IAGRSSA does not require any decision column on the dataset, and yet its computational complexity is far below that of the direct algorithms. Ideally, the block construction process should involve supervision as it requires an understanding of the attributes. AGRSSA-M [29] is based on reducing the blocks required for decisions.

Theorem 9. *IAGRSSA computes the blocks of the tolerance constructed as a direct product of the tolerances on each column.*

Proof. First, chains and partial orders on a set are equivalent to specific groupoidal operations [9], [34], and the compatibility is assumed with respect to these. Additionally, direct products

of groupoids are groupoids. The rest follows from Theorems 5 and 3. □

The next example illustrates the product.

Example 3. *Let T_1 be the tolerance on Q defined by*

$$T_1 ab \text{ if and only if } |a-b| \leq 1,$$

and T_2 be the tolerance on $\Re$ defined by

$$T_2 ab \text{ if and only if } |a-b| \leq e.$$

On the product set $Q \times \Re$ with the induced lattice order, the product tolerance T is defined by the condition

$$T(a_1, a_2)(b_1, b_2) \text{ if and only if } T_1 a_1 b_1 \& T_2 a_2 b_2.$$

The blocks of the tolerance T_1 are of the form $\{x : |x-q| \leq 0.5\}$ for distinct $q \in Q$ The blocks of the tolerance T_2 on the other hand are of the form $\{x : |x-a| \leq 0.5e\}$ for distinct $a \in \Re$. The blocks of the product need to be formed by taking a direct product of these as the components are independent. It is therefore the set

$$\{(x, w) : |x-q| \leq 0.5 \& |w-a| \leq 0.5e\} \text{ for } q \in Q \,\&\, a \in \Re.$$

Exhaustive Tolerance Discovery Algorithm (ETDA-LMR)

A *large-minded reasoner* is so-named because it is essentially about discovering suitable similarities. It selects the more reasonable collections of blocks. The exhaustive tolerance discovery algorithm that involves LMRs is invented in the paper [29]. However, it is relatively opaque as it leaves out the crucial steps of selection to the dynamics of the context. A natural question is: can the simplicity of the structure of blocks on chains be exploited to improve the meta algorithm.

Definition 7. *A* large-minded reasoner *(LMR) is a partial function* $\psi : \mathbf{UBD}(A_1) \times \mathbf{UBD}(A_2) \times \ldots \mathbf{UBD}(A_n) \longmapsto \mathbf{UBD}(A)$.

The ETDA algorithm applies to all kinds of information tables including decision tables, and in clearer terms is given below.

EDTA Algorithm:

Step 1: Define sequences of q number of quantiles – this by itself means a certain understanding of categories associated with attributes.
Step 2: Using the quantiles form intervals (with or without intersections) under the conditions of Theorem 3.
Step 3: Optionally, some intervals may be fused together in relation to relative changes in decisions. This amounts to removing interval boundaries.
Step 4: Specify the large minded reasoner ψ. This is the same as defining a number of compatible tolerances using the intervals.
Step 5: Identify the defined tolerances in ψ.
Step 6: Compute relevant lower, upper, and bited approximations and optionally the associated decisions for each normal cover.
Step 7: In case of soft cluster validation, compute the approximations of the cores and exteriors, and evaluate their closeness to the evaluated.
Step 8: Select relevant tolerances (or normal covers) specified in ψ.
Step 9: Explain the data context on the basis of the associated tolerance(s) (or normal covers).

Example 4. *The first three steps of the ETDA algorithm are illustrated in this example. Let $\{1, 5, 6, 9, 10\}$ be a sequence of equally spaced quantiles. Some sets of intervals that can be formed with these are*

$$\mathcal{B}_1 = \{[1,6],\ [5,9], [6,10]\}$$
$$\mathcal{B}_2 = \{[1,5],\ [5,6], [6,9], [9,10]\}$$
$$\mathcal{B}_3 = \{[1,9],\ [5,10]\}.$$

Definition 8. *By an* interpreted large-minded reasoner *associated with ψ of Def. 7 will be meant a partial function $\psi^* : \mathbf{UB}(A_1) \times \mathbf{UB}(A_2) \times \ldots \mathbf{UB}(A_n) \times \wp(S) \longmapsto \mathbf{UB}(A)$, that indicates the granular components or parts of approximations of subsets.*

The function is intended to represent the compositionality of approximations in terms of blocks of components. These can be quite complex (see [5], [9]), and so *granular components or parts of approximations* are referred to.

Theorem 10. *ψ^* is a CRRF of type H.*

VI. Problems and Directions

The invented algorithms appear to be well-suited for low cost computing. A detailed investigation is necessary to confirm the same. Additionally, it is necessary to formulate post-processing techniques for seamless interpretation. If a context results in a hundred blocks, then a description of the blocks, and the approximations generated is essential for keeping track of meaning. How does one solve this *problem of meaningful empirical representation*?

One way is to encode the blocks lexicographically on the basis of its position on components (or columns). Next, the extent of expression of these blocks (encoding) in relatively important clusters can be computed. The combination of these expressions can be expressed in natural language with limited or no involvement of numeric estimates. A full solution of this problem will appear separately.

The logic of decision-making on the basis of set-theoretic measures for cluster validation requires additional work and will appear in a forthcoming paper. A substantial amount of the machinery required is invented in the papers [35], and a forthcoming three part paper by the first author.

In similarity-based clustering [36], clusters are formed from data supplemented with similarity grades (usually with values in the real interval $[0, 1]$ or the rationals Q) between data points. For a set of n data points, the associated similarity matrix formed by these similarity grades is a symmetric square matrix $K = (s_{ij})_{n \times n}$ with s_{ij} being the similarity between the ith and jth data point. These can as well be approached

through spectral clustering methods. Can the EDTA algorithm be extended to these contexts?

A. Problems of Medical Imaging and Beyond

Big datasets associated with medical images (obtained through MRI, FMRI and CT scans) are mostly patterns formed by products of finite totally ordered subsets of the reals. Rough sets combined with clustering techniques are used to identify brain tumors in the presence of bias field and noise in recent work [37]. However, additional methods need to be employed to possibly rectify the results. Specifically, the CoLoRS segmentation algorithm does not clearly provide the reasons for inclusion or exclusion of tumors or healthy tissues. It is of interest to use the transparent algorithms invented in this research to these problem contexts, and additionally in those for identification of lesions in lung CT [38]. These characteristics are typical of a number of other application contexts of AIML, and therefore the application contexts are boundless.

REFERENCES

[1] C. Hennig, M. Meila, F. Murtagh, and R. Rocci, *Handbook of Cluster Analysis*, 1st ed. Edited Volume. Chapman and Hall: CRC Press, 2016.

[2] M. Kim and R. S. Ramakrishna, "New Indices For Cluster Validity Assessment Pattern," *Pattern Recognition Letters*, vol. 26, pp. 2353–2363, 2005. doi: 10.1016/j.patrec.2005.04.007

[3] C. Hennig and T. F. Liao, "Comparing Latent Class and Dissimilarity Based Clustering for Mixed Type Variables..." *Journal of the Royal Statistical Society*, pp. 309–369, 2013. doi: 10.1111/j.1467-9876.2012.01066.x

[4] C. Bouveyron, G. Celeux, B. Murphy, and A. Raftery, *Model-Based Clustering and Classification for Data Science: With Applications in R*. Cambridge University Press, 2019.

[5] A. Mani, "Dialectics of Counting and The Mathematics of Vagueness," *Transactions on Rough Sets*, vol. XV, no. LNCS 7255, pp. 122–180, 2012. doi: 10.1007/978-3-642-31903-7_4

[6] ——, "Comparative Approaches to Granularity in General Rough Sets," in *IJCRS 2020*, ser. LNAI, R. Bello *et al.*, Eds. Springer, 2020, vol. 12179, pp. 500–518.

[7] ——, "General Rough Modeling of Cluster Analysis," in *Rough Sets: IJCRS-EUSFLAT 2021*, ser. LNAI 12872, S. Ramanna *et al.*, Eds. Springer Nature, 2021.

[8] A. Campagner and D. Ciucci, "A formal learning theory for three-way clustering," in *SUM 2020*, J. Davis and K. Tabia, Eds. Springer, 2020, vol. LNAI 12322, pp. 128–140.

[9] A. Mani, "Algebraic Methods for Granular Rough Sets," in *Algebraic Methods in General Rough Sets*, ser. Trends in Mathematics, A. Mani, I. Düntsch, and G. Cattaneo, Eds. Birkhauser Basel, 2018, pp. 157–336.

[10] ——, "Algebraic Semantics of Similarity-Based Bitten Rough Set Theory," *Fundamenta Informaticae*, vol. 97, no. 1-2, pp. 177–197, 2009. doi: 10.3233/FI-2009-196

[11] P. Wasilewski and D. Ślęzak, "Foundations of Rough Sets from Vagueness Perspective," in *Rough Computing: Theories, Technologies and Applications*, A. Hassanien *et al.*, Eds. IGI, Global, 2008, pp. 1–37.

[12] I. Chajda, *Algebraic Theory of Tolerance Relations*. Olomouc University Press, 1991. [Online]. Available: https://www.researchgate.net/publication/36797871

[13] A. Mani, "Representation, Duality and Beyond," in *Algebraic Methods in General Rough Sets*, ser. Trends in Mathematics, A. Mani, I. Düntsch, and G. Cattaneo, Eds. Birkhauser Basel, 2018, pp. 459–552.

[14] H. J. Bandelt, "Tolerance relations of a lattice," *Bulletin Austral.Math. Soc.*, vol. 23, pp. 367–381, 1981. doi: 10.1017/S0004972700007255

[15] G. Czedli and L. Klukovits, "A note on tolerances of idempotent algebras," *Glasnik Matematicki (Zagreb)*, vol. 18, pp. 35–38, 1983. [Online]. Available: https://web.math.pmf.unizg.hr/glasnik/18.1/18103.pdf

[16] A. Górnicka, J. Grygiel, and I. Tyrala, "On the lattice of tolerances for a finite chain," *Czestochowa Mathematics*, vol. XXI, pp. 25–30, 2016. doi: 10.16926/m.2016.21.03

[17] J. Niederle, "A note on tolerance lattices of products of lattices," *Casop. Pest. Matem.*, vol. 107, pp. 114–115, 1982. doi: 10.21136/CPM.1982.118112

[18] S. Mitra, "An Evolutionary Rough Partitive Clustering," *Pattern Recognition Letters*, vol. 25, no. 12, pp. 1439–1449, 2004. doi: 10.1016/j.patrec.2004.05.007

[19] G. Peters, "Rough Clustering Utilizing the Principle of Indifference," *Information Sciences*, vol. 277, pp. 358–374, 2014. doi: 10.1016/j.ins.2014.02.073

[20] J. Vermunt and J. Magidson, "Latent Class Cluster Analysis," in *Applied Latent Class Analysis*. Cambridge: Cambridge University Press, 2002, pp. 89–106.

[21] I. Düntsch and G. Gediga, "Rough Set Clustering," in *Handbook of Cluster Analysis*, C. Hennig, M. Meila, and F. Murtagh, Eds. CRC Press, 2016, ch. 25, pp. 575–594.

[22] A. D. Concilio, C. Guadagni, J. Peters, and S. Ramanna, " Descriptive Proximities. Properties and Interplay Between Classical Proximities and Overlap," *Mathematics in Computer Science*, vol. 12, no. 1, pp. 91–106, 2018. doi: 10.1007/s11786-017-0328-y

[23] A. N. Kolmogorov, "On the logical foundations of probability theory," in *Selected Works of A. N. Kolmogorov*, A. N. Shiryayev, Ed. Kluwer Academic, Nauka, 1986, vol. 2, ch. 53, pp. 515–519.

[24] B. Liu, *Uncertainty Theory*, ser. Studies in Fuzziness and Soft Computing. Springer, 2004, vol. 154.

[25] T. Steifer, "A note on learning theoretic characterizations of randomness and convergence," *Review of Symbolic Logic*, vol. 15, no. 3, pp. 807–822, 2022. doi: 10.1017/S1755020321000125

[26] L. Beach and G. Braun, "Laboratory studies of subjective probability: a status report," in *Subjective Probability*, G. Wright and P. Ayton, Eds. John Wiley, 1994, pp. 107–128.

[27] A. Mani, I. Düntsch, and G. Cattaneo, Eds., *Algebraic Methods in General Rough Sets*, ser. Trends in Mathematics. Birkhauser Basel, 2018. ISBN 978-3-030-01161-1

[28] P. Pagliani and M. Chakraborty, *A Geometry of Approximation: Rough Set Theory: Logic, Algebra and Topology of Conceptual Patterns*. Berlin: Springer, 2008.

[29] A. Mani, "Rough randomness and its application," *Journal of the Calcutta Mathematical Society*, pp. 1–15, 2023. doi: 10.5281/zenodo.7762335. [Online]. Available: https://zenodo.org/record/7762335

[30] J. Stepaniuk, *Rough-Granular Computing in Knowledge Discovery and Data Mining*, ser. Studies in Computational Intelligence,Volume 152. Springer-Verlag, 2009. ISBN 978-3-540-70800-1

[31] A. Gomolinska, "Rough Approximation Based on Weak q-RIFs," *Transactions on Rough Sets*, vol. X, pp. 117–135, 2009. doi: 10.1007/978-3-642-03281-3_4

[32] M. Gagrat and S. Naimpally, "Proximity approach to semi-metric and developable spaces," *Pacific Journal of Mathematics*, vol. 44, no. 1, pp. 93–105, 1973. doi: 10.2140/pjm.1973.44-1

[33] O. Cheong *et al.*, Eds., *Listing All Maximal Cliques in Sparse Graphs in Near-Optimal Time*, ser. LNCS 6506. Berlin: Springer, 2010. doi: 10.1007/978-3-642-17517-6_36

[34] R. Freese, J. Jezek, J. Jipsen, P. Markovic, M. Maroti, and R. Mckenzie, "The Variety Generated by Order Algebras," *Algebra Universalis*, vol. 47, pp. 103–138, 2002. doi: 10.1007/s00012-002-8178-z

[35] A. Mani, "Granularity and Rational Approximation: Rethinking Graded Rough Sets," *Transactions on Rough Sets*, vol. XXIII, no. LNCS 13610, pp. 33–59, 2022. doi: 10.1007/978-3-662-66544-2_4

[36] M. Meila, "Spectral Clustering," in *Handbook of Cluster Analysis*, C. Hennig, M. Meila, and F. Murtagh, Eds. CRC Press, 2016, ch. 7, pp. 125–144.

[37] S. Roy and P. Maji, "Tumor delineation from 3-d mr brain images," *Signal, Image and Video Processing*, pp. 1–9, 2023. doi: 10.1007/s11760-023-02565-4

[38] P. Dutta and S. Mitra, "Composite deep network with feature weighting for improved delineation of covid infection in lung ct," *MedRxiv*, pp. 1–22, 2023. doi: 10.1101/2023.01.17.23284673

Position Papers of the 18th Conference on Computer Science and Intelligence Systems pp. 9–16
DOI: 10.15439/2023F5821
ISSN 2300-5963 ACSIS, Vol. 36

Key Factors Affecting Knowledge Sharing in Developed and Developing Countries: a specific focus on Saudi Arabia

Abdulrahman Alsaif
0009-0003-6173-661X
Department of Software Engineering, Prince Sattam University, Saudi Arabia
Department of Informatics, University of Sussex, United Kingdom
aa2716@sussex.ac.uk

Imran Khan
0000-0002-5732-4685
Department of Informatics, University of Sussex, United Kingdom
Imran.khan@sussex.ac.uk

Martin White
0000-0001-8686-2274
Department of Informatics, University of Sussex, United Kingdom
m.white@sussex.ac.uk

Natalia Beloff
0000-0002-8872-7786
Department of Informatics, University of Sussex, United Kingdom
n.beloff@sussex.ac.uk

***Abstract*—Knowledge is considered the most valuable asset in the modern digital economy, and its dissemination is recognised as the backbone of successful enonomies. The dissemination of knowledge is widely recognised as a fundamental practise in any successful organization, including higher education institutes. This practise enables institutions to generate and maintain knowledge. Organizations that foster a culture of knowledge sharing are able to gain a competitive advantage and drive innovation. These organisations contribute to the enhancement of the economy. The Kingdom of Saudi Arabia has implemented strategies to incorporate knowledge-sharing initiatives within its organisations. The primary objective of this paper is to identify the key factors that impact knowledge sharing in both developed and developing nations, with a particular focus on Saudi Arabia, and to present our future work on how to deploy cloud computing for knowledge sharing in Saudi Arabia's HEIs.**

***Index Terms*—knowledge sharing, developed/developing countries, Saudi Arabia.**

I. Introduction

IN TODAY'S interconnected world, data, information, and knowledge play an increasingly important role in the development and progression of civilizations. Information is the form of data that has been organized, interpreted, and given meaning. As opposed to mere facts, knowledge incorporates comprehension, insights, and the ability to utilise information effectively [7]. Knowledge is essential for empowering individuals, organizations, and nations to make prudent decisions, foster innovation, and advance society [1]–[3]. It serves as a catalyst for social progress, economic growth, and general prosperity.

Knowledge is seen as an advantage in both developed and developing countries [4], [5]. The developed countries have institutions and programmes in place to facilitate collaboration, information sharing, and knowledge [55]. It is important to highlight factors of knowledge sharing that can be adopted by organisations in developing nations like Saudi Arabia. Saudi Arabia is among the fastest-developing nations. The core of its ambitious vision 2030 is to develop infrastructure to reduce dependence on an oil economy and move into a knowledge economy. In this regard, it is important for Saudi Arabia to nurture an environment and culture of knowledge sharing. Both public and private organisations encourage their employees to share their knowledge and expertise [59]. However, an understanding of knowledge sharing from the Saudi perspective is lacking. Thus, the research paper conducts a systematic literature review to identify the key factors associated with knowledge sharing in developing and developed countries.

II. Methodology

To answer the above research aim, this paper adopted a systematic literature review to highlight the significant factors of knowledge sharing. First, we defined our main research question, which is:

What are the key factors associated with successful knowledge-based economies?

To answer these research questions, we studied the literature focused on four key areas:

- Data versuse Information versus Knowledge
- Knowledge sharing
- Knowledge sharing in developed/developing countries
- Knowledge sharing in Saudi Arabia

This is discussed in more detail in the associated sections below. In each of these sections, we implement tables that summarise the key factors discovered in the literature review.

III. Data vs Information vs Knowledge

Knowledge is the most valuable asset of any industrial business or academic institution [6]. The information era was recognised by the growth of information technology and the digital era due to the development of technology, particularly the Internet [7]. Some scholars [8]–[10] have long maintained

the view that knowledge is power. Knowledge comes through interpreting information, and that information comes from the attribution of meaning to data [11], [12].

Despite numerous definitions, it seems that there is still a lack of clarity around what data, information, and knowledge are and the way they relate to one another [12], [13]. Based on the Critical Delphi project [12], which used a qualitative research approach to facilitate dialogues among 57 experts from 16 different nations. They came to the conclusion that knowledge, information, and data all have a certain order. Data can only be used to generate information, and knowledge can only be used to create information [11], [12]. During their studies, they rely on the definitions of data, information, and knowledge rather than other notions such as correctness, adequacy, and definition coherence. Based on [12], Table I and Appendix A show some of the key definitions of data, information, and knowledge.

TABLE I.
KEY DEFINITION OF DATA, INFORMATION AND KNOWLEDGE

Data	raw material of information, facts, symbols, and basic individual items, numeric, unprocessed, eligible to be processed to produce knowledge, without context and interpretation.
Information	has meaning, is able to be analysed and interpreted, has purpose, has been communicated, has been categorized, and has the ability to create knowledge.
Knowledge	Accumulated information, tempered by experience, meaningful, information with more context, make a difference in an enterprises, emerges from analysis.

The systematic literature study helps to conclude that data are the basic forms or fundamental units of numbers, characters, symbols, and signal readings, as well as other information such as audio, video, and text, that have been acquired by observation but are meaningless on their own. While information is described as facts, implication meaning, input, and other sorts of meaningful representations that, when encountered or provided to a human person, are used to increase his or her understanding of a subject or associated concerns, aid in decision-making, or solve problems, On the other hand, knowledge [14] is the capacity to act (know-how), recognise (know-what), and comprehend (know-why), and it is anything that exists or is stored in the mind or brain in order to better our lives and add value.

IV. KNOWLEDGE SHARING

According to Peter Drucker [15], knowledge rather than money, capital, or even technology forms the basis of the twenty-first century corporation. The knowledge has to be shared in order to produce value [15]–[17]. The definition of "knowledge sharing" has been attempted several times [16], [18], and [19], but depending on the context and point of view, it is still widely discussed among academics and practitioners [19]. An analysis of many knowledge sharing definitions according to [16], [18], and [19] shows a commonality in their wording that it is an activity that includes the interchange of information and knowledge across people, businesses, and communities. There are two basic types of knowledge sharing: explicit and tacit [20], [21]. According to [20], explicit knowledge could be expressed verbally, written down, represented numerically, or represented visually. Contrarily, tacit knowledge is defined as "information that cannot be readily expressed in words and is not readily comprehensible."

Some scholars [1], [16], and [17] found that sharing information inside businesses had several advantages. It may enhance decision-making, reduce redundancies, boost innovation, and make operations more efficient. Additionally, it enables the exchange of best practices, which may help firms retain their competitiveness by upholding the highest levels of quality. Additionally, information sharing among staff members in a company may foster a sense of teamwork, which can boost morale and increase job satisfaction. In general, information sharing is a critical element for the success of organizations.

V. KNOWLEDGE SHARING IN DEVELOPED/DEVELOPING COUNTRIES

Knowledge sharing is becoming highly significant in developed countries as organisations attempt to remain competitive and encourage innovation [4], [5]. In developed countries, knowledge can be shared through various techniques, including conferences, webinars, formal and informal networks, and other digital platform technologies [7]. whereas it is well known that businesses situated in some developing countries lack managerial acumen skills, technical tools, and other financial resources [22], [23].

Knowledge sharing in developed or developing countries may be affected by various elements either positively or negatively [16], [23], [24]. Many factors contributing to knowledge sharing behaviour have been recognised and clarified by various researchers, for example, top management support [25], [26], individual willingness [27], reward systems and motivation [28], and information technology adoption [29], including emails, websites, and online discussion forums [30]. Companies in the US use collaboration technologies like Slack and Zoom to encourage knowledge exchange among their staff members.

Professionals can communicate with one another and discuss best practises through professional networks like LinkedIn. Universities and other educational institutions frequently organise seminars and workshops to exchange information with their faculty and students.

Conversely, there is hardly much knowledge exchange in poor nations [31]. A number of reasons were reported, such as inadequate infrastructure, restricted access to technology [22], [23], and a dearth of institutions for education and research. As a result, there is frequently a restriction on the flow of knowledge between people and organizations. The capacity of people and organisations to communicate knowledge is also constrained by a lack of technology and resources. For instance, it can be challenging for people to exchange knowledge and work together when there is little access to the internet and other kinds of communication in some rural regions [31]. Studying the literature has helped to provide Table II, which lists the most common factors of knowledge sharing in developed and developing countries.

TABLE II.

COMMON FACTORS IN DEVELOPED/DEVELOPING COUNTRIES

Factor	Country	Author
Trust, security and privacy concerns	USA, China, Taiwan, Sweden,	[32], [33]
Openness to change	Saudi Arabia, Brazil	[34], [35]
Individual's willingness	USA, Dutch, Europe, Asia, Australia	[32], [36], [37]
Information Technology & social media	China, Canada & Australia, Iran, Taiwan, Hong Kong	[38]–[43]
Resource constraints	Australia, Sub-Saharan Africa	[44]–[46]
cultural norms, Organisational culture	Finland, Saudi Arabia	[34], [47]
Top management	Italy	[25], [48]
organizational commitment	China	[49]–[51]
Incentives and reward system	Finland, Taiwan, Malaysia	[47], [52], [53]

Numerous actions are being taken to encourage information exchange in both developed and developing countries. A programme called "knowledge sharing" has been established by the National Science Foundation [54] in the US to support research initiatives that foster information exchange between US scientists and their international colleagues. The initiative offers funding for global partnerships, workshops, and conferences that advance international scientific knowledge exchange. The NSF has also provided funding for research projects that examine how technology might be used to improve knowledge exchange in developed nations. For instance, the NSF provided funding for a project that created a platform for scientists to work together on research initiatives and successfully communicate their findings to other scientists.

A multinational network of people and organisations known as the Knowledge Sharing Alliance (KSA) has gathered to exchange resources and best practises in order to promote knowledge-based economies. For the purpose of fostering a more dynamic and competitive global economy, KSA encourages its members to share and exchange knowledge, concepts, and experiences. As an illustration, KSA members can work together to create novel solutions to problems affecting the knowledge economy between nations with advanced technological infrastructure and those without it. KSA also gives users and governments access to a platform where they can get current data and resources on subjects relating to knowledge sharing, including water security, policy frameworks for investment, urban developments, and others [55].

VI. KNOWLEDGE SHARING IN SAUDI ARABIA

Saudi Arabia has the largest economy in the Middle East and the Arab world and is among the top twenty economies in the world. It is dominated by the oil industry, which generates around 87% of budgetary income, 90% of export revenue, and 42% of GDP [56]. The greatest petroleum exporter in the world and a superpower in the energy sector is the Kingdom of Saudi Arabia [57], [58].

To expand the nation's ability to produce goods and conduct business, the government has built a number of economic and industrial cities [59]. Additionally, it has created a number of free trade zones that are intended to promote international investment. Saudi Arabia is a popular location for international direct investment, and the government has worked hard to draw in outside capital.

Knowledge sharing in Saudi Arabia is becoming increasingly important as the nation strives to become a global leader in various industries, including technology and energy. To facilitate knowledge sharing, the government has implemented several initiatives to promote collaboration, such as creating a legal framework to protect intellectual property rights and encouraging the formation of research and development centers. Additionally, universities and educational institutions are trying to develop innovative methods to facilitate the exchange [60].

On June 30, 2009, the Saudi Corporation for Electronic Information Exchange [61] was created with the goal of investing in communication and information technology

projects as well as knowledge-based businesses. The company's business field is to develop apps and e-transaction solutions for a new concept of electronic data interchange and transmission between the Customs Authority and other pertinent private and public entities in the exports and imports industry that have never been covered. Tabadul has created the Fasah platform, which simplifies import and export processes and provides a range of services that promote global trade by connecting with the appropriate authorities, tracking shipments, scheduling appointments, and offering online payment options.

Several variables, including the availability of technology [62], [63], cultural values [17], [58], the desire of individuals to share knowledge [58], and the availability of trustworthy and accurate information [64], [65], can either promote or inhibit knowledge sharing in Saudi Arabia. Many studies about factors that enable or hinder knowledge sharing in Saudi Arabia from different sectors have been reviewed; see Table III.

TABLE III.

COMMON FACTORS THAT AFFECT KNOWLEDGE SHARING IN SAUDI ARABIA

Factor	Sectors	Author
Demographics	Private companies, of varied size	[58]
Openness in communication	Education, Saudi Arabian organisations	[17], [65]
Interpersonal trust, trust, privacy	Educationm, Health	[64], [65]
Perceived usefulness and perceived ease of use	Education, Health Information	[64], [65]
Motivation and reward system	Education, Saudi Arabia Organization, eLearning Virtual Communities, Saudi Telecom STC	[17], [34], [63], [65]
Management support	Telecommunications	[17]
Nature of knowledge	Universities	[60]
Information and communication technology	Industrial and commercial sectors	[62], [63]

VII. CONCLUSIONS

This research concludes and highlights that knowledge is an important asset for developing the economies of developing nations like Saudi Arabia. The paper provides clear definitions of important terms: data, information, knowledge, and knowledge sharing (see Table I). The paper has carried out a systematic investigation of the literature to highlight the factors that impact knowledge sharing in developing and developed nations. Table II provides a list of these knowledge-sharing factors. The paper has also contributed a list of factors that are considered significant to the sharing of knowledge in Saudi Arabia; see Table III.

Overall, Sections III, IV, V, and VI articulated the differences and similarities between existing literature on knowledge sharing in developed and developing economies. Practitioners of knowledge management can use the factors outlined in Tables I, II, and III to assess their knowledge sharing plans. In our case, we plan to build a new knowledge-sharing cloud-based platform.

While this research paper answers our basic research question, "What are the key factors associated with successful knowledge-based economies?" Tables II and III contrast factors within Saudi Arabia as a developing nation and other developed and developing economies. It is important to note that the research paper has limitations. The factors highlighted have strong roots in the literature, which needs to be validated by building a new framework based on Knowledge Sharing Software-Based Cloud (KSSbC); see future work.

VIII. FUTURE WORK

According to [20], [66], knowledge is currently considered to be one of an organization's key assets, alongside labor, land, and money, since it gives organisations a competitive advantage. Organizations have recognised the value of information and the benefits of managing it well, including enhancing performance, boosting productivity, and increasing profitability [1]–[3]. However, Arab countries, especially Saudi Arabia, are failing to share their knowledge. Furthermore, some scholars believe that information exchange inside Saudi Arabia's Higher Education Institute is required [67].

Due to the characteristics of cloud computing, numerous universities expressed interest in integrating cloud computing into their educational systems [68]. In order for HEIs to adopt cloud services, a clear cloud strategy that supports CC capabilities is necessary. In order to implement the cloud services plan, a new framework must be developed that meets the needs of key stakeholders, including academics, students, and HEI board directors. To have a successful cloud strategy, the key stakeholder should be involved in defining the HEI's cloud strategy, which tackles its opportunities, problems, and concerns specific to HEIs, as well as the need for the cloud strategy to be in line with the HEI's plan [69].

The research aims to enhance knowledge sharing practises among beneficiaries in Saudi universities by addressing the gap between knowledge sharing contexts and cloud computing. Therefore, we are intending to design a prototype platform named Knowledge Sharing Software-based Cloud (KSSbC) using software engineering methodologies. Based on that, a new framework will be built, tested, and validated, with key factors affecting it either positively or negatively.

APPENDIX

APPENDIX A.

DATA, INFORMATION, AND KNOWLEDGE DEFINITIONS.

Author	Definition
[70]	**Data:** "Data are the basic individual items of numeric or other information, garnered through observation; but in themselves, without context, they are devoid of information." **Information:** "Information is that which is conveyed, and possibly amenable to analysis and interpretation, through data and the context in which the data are assembled." **Knowledge:** "Knowledge is the general understanding and awareness garnered from accumulated information, tempered by experience, enabling new contexts to be envisaged."
[71]	**Data:** "Data are a string of symbols." **Information:** "Information is data that is communicated, has meaning, has an effect, has a goal." **Knowledge:** "Knowledge is a personal/cognitive framework that makes it possible for humans to use information."
[72]	**Data:** "Data are the raw observations about the world collected by scientists and others, with a minimum of contextual interpretation. " **Information:** "Information is the aggregation of data to make coherent observations about the world. " **Knowledge:** "Knowledge is the rules and organizing principles gleaned from data to aggregate it into information. "
[73]	**Data:** "Data are raw material of information, typically numeric. " **Information:** "Information is data which is collected together with commentary, context and analysis so as to be meaningful to others. " **Knowledge:** "Knowledge is a combination of information and a person's experience, intuition and expertise. "
[74]	**Data:** "Data are raw evidence, unprocessed, eligible to be processed to produce knowledge." **Information:** "Information is the process of becoming informed; it is dependent on knowledge, which is processed data. Knowledge perceived, becomes information. " **Knowledge:** "Knowledge is what is known, more than data, but not yet information. Recorded knowledge may be accessed in formal ways. Unrecorded knowledge is accessible in only chaotic ways. "
[75]	**Data:** "Data are representations of facts and raw material of information. " **Information:** "Information is data organized to produce meaning. " **Knowledge:** "Knowledge is meaningful content assimilated for use. The three entities can be viewed as hierarchical in terms of complexity, data being the simplest and knowledge, the most complex of the three. Knowledge is the product of a synthesis in our mind that can be conveyed by information, as one of many forms of its externalization and socialization. "
[76]	**Data:** "Data are facts and statistics that can be quantified, measured, counted, and stored. " **Information:** "Information is data that has been categorized, counted, and thus given meaning, relevance, or purpose. " **Knowledge:** "Knowledge is information that has been given meaning and taken to a higher level. Knowledge emerges from analysis, reflection upon, and synthesis of information. It is used to make a difference in an enterprise, learn a lesson, or solve a problem. "
[77]	**Data:** "Data are atomic facts, basic elements of "truth," without interpretation or greater context. It is related to things we sense. " **Information:** "Information is a set of facts with processing capability added, such as context, relationships to other facts about the same or related objects, implying an increased usefulness. Information provides meaning to data. " **Knowledge:** "Knowledge is information with more context and understanding, perhaps with the addition of rules to extend definitions and allow inference. "
[78]	**Data:** "Data is a symbol set that is quantified and/or qualified. " **Information:** "Information is a set of significant signs that has the ability to create knowledge . . . The essence of the information phenomenon has been characterized as the occurrence of a communication process that takes place between the sender and the recipient of the message. Thus, the various concepts of information tend to concentrate on the origin and the end point of this communication process. " **Knowledge:** "Knowledge is information that has been appropriate by the user. When information is adequately assimilated, it produces knowledge, modifies the individual's mental store of information and benefits his development and that of the society in which he lives. Thus, as the mediating agent in the production of knowledge, the information, qualifies itself, in form and substance, as significant structures able to generate knowledge for the individual and his group. "

REFERENCES

[1] N. Bontis, "Assessing knowledge assets: a review of the models used to measure intellectual capital," International journal of management reviews, vol. 3, no. 1, pp. 41–60, 2001. https://doi.org/10.1111/1468-2370.00053

[2] A. Ismail Al-Alawi, N. Yousif Al-Marzooqi, and Y. Fraidoon Mohammed, "Organizational culture and knowledge sharing: critical success factors," *Journal* of knowledge management, vol. 11, no. 2, pp. 22–42, 2007. https://doi.org/10.1108/13673270710738898

[3] H. Hussinki, A. Kianto, M. Vanhala, and P. Ritala, "Assessing the universality of knowledge management practices," Journal of Knowledge Management, vol. 21, no. 6, pp. 1596–1621, 2017. https://doi.org/10.1108/JKM-09-2016-0394

[4] M. Mohsin and J. Syed, "Knowledge management in developing economies: A critical review," *The Palgrave Handbook of Knowledge Management*, pp. 601–620, 2018.

[5] A. Y. Noaman and F. Fouad, "Knowledge sharing in universal societies of some develop nations," *Int J Acad Res*, vol. 6, no. 3, pp. 205–212, 2014.

[6] H. Zhuge, "A knowledge flow model for peer-to-peer team knowledge sharing and management," *Expert* Syst Appl, vol. 23, no. 1, pp. 23–30, 2002. https://doi.org/10.1016/S0957-4174(02)00024-6

[7] Z. Gaál, L. Szabó, N. Obermayer-Kovács, and A. Csepregi, "Exploring the role of social media in knowledge sharing," *Electronic Journal of Knowledge Management*, vol. 13, no. 3, pp. pp185-197, 2015.

[8] M.-Y. Cheng, J. S.-Y. Ho, and P. M. Lau, "Knowledge sharing in academic institutions: A study of Multimedia University Malaysia.," *Electronic Journal of knowledge management*, vol. 7, no. 3, 2009.

[9] M. J. Iqbal, A. Rasli, L. H. Heng, M. B. B. Ali, I. Hassan, and A. Jolaee, "Academic staff knowledge sharing intentions and university innovation capability," *African Journal of Business Management*, vol. 5, no. 27, p. 11051, 2011. https://doi.org/10.5897/AJBM11.576

[10] K. K. Jain, M. S. Sandhu, and G. K. Sidhu, "Knowledge sharing among academic staff: A case study of business schools in Klang Valley, Malaysia." UCSI Centre for Research Excellence, 2007.

[11] *G. Jifa, "Data, information, knowledge, wisdom and meta-synthesis of wisdom-comment on wisdom global and wisdom cities," Procedia Comput Sci, vol. 17, pp. 713–719,2013. https://doi.org/10.1016/j.procs.2013.05.092*

[12] C. Zins, "Conceptual approaches for defining data, information, and knowledge," Journal of the American society for information science and technology, vol. 58, no. 4, pp. 479–493, 2007. https://doi.org/10.1002/asi.20508

[13] "Def." http://vlibrary.info/InfoLexicon.html (accessed Mar. 04, 2023).

[14] R. Garud, "On the distinction between know-how, know-why, and know-what," *Advances in Strategic Management*, vol. 14, pp. 81–101, Jan. 1997.

[15] C. M. Jacobson, "Knowledge sharing between individuals," in *Encyclopedia of Knowledge Management, Second Edition*, IGI Global, 2011, pp. 924–934.

[16] M. Asrar-ul-Haq and S. Anwar, "A systematic review of knowledge management and knowledge sharing: Trends, issues, and challenges," *Cogent Business &* Management, vol. 3, no. 1, p. 1127744, 2016. https://doi.org/10.1080/23311975.2015.1127744

[17] R. K. Yeo and J. Gold, "Knowledge sharing attitude and behaviour in Saudi Arabian organisations: why trust matters," *International Journal of Human Resources Development and Management*, vol. 14, no. 1–3, pp. 97–118, 2014.

[18] M. Ipe, "Knowledge sharing in organizations: A conceptual framework," *Human resource development review, vol. 2, no. 4, pp. 337–359, 2003. https://doi.org/10.1177/1534484303257985*

[19] C. Nielsen and K. Cappelen, "Exploring the mechanisms of knowledge transfer in University-Industry collaborations: A study of companies, students and researchers," *Higher Education* Quarterly, vol. 68, no. 4, pp. 375–393, 2014. https://doi.org/10.1111/hequ.12035

[20] M. Alavi and D. E. Leidner, "Knowledge management and knowledge management systems: Conceptual foundations and research issues," MIS quarterly, pp. 107–136, 2001. https://doi.org/10.2307/3250961

[21] *A. Fengjie, Q. Fei, and C. Xin, "Knowledge sharing and web-based knowledge-sharing platform," in IEEE International Conference on E-commerce Technology for Dynamic E-business, IEEE, 2004, pp. 278–281. https://doi.org/10.1109/CEC-EAST.2004.43*

[22] J. Kuada, "Collaboration between developed and developing country-based firms: Danish-Ghanaian experience," Journal of Business & Industrial Marketing, vol. 17, no. 6, pp. 538–557, 2002. https://doi.org/10.1108/08858620210442866

[23] M. I. Manda and J. Backhouse, "An analysis of the barriers to e-government integration, interoperability and information sharing in developing countries: A systematic review of literature," in *Proceedings of the African Conference in Information Systems and Technology, Accra, Ghana*, 2016, pp. 5–6.

[24] H. Alotaibi, R. Crowder, and G. Wills, "Investigating factors for E-knowledge sharing amongst academic staff," 2014.

[25] V. Cavaliere and S. Lombardi, "Exploring different cultural configurations: how do they affect subsidiaries' knowledge sharing behaviors?," Journal of Knowledge Management, vol. 19, no. 2, pp. 141–163, 2015. https://doi.org/10.1108/JKM-04-2014-0167

[26] R. Farooq, "A conceptual model of knowledge sharing," International Journal of Innovation Science, 2018. https://doi.org/10.1108/IJIS-09-2017-0087

[27] B. Van Den Hooff, A. P. Schouten, and S. Simonovski, "What one feels and what one knows: the influence of emotions on attitudes and intentions towards knowledge sharing," Journal of knowledge management, vol. 16, no. 1, pp. 148–158, 2012. https://doi.org/10.1108/13673271211198990

[28] G. Tangaraja, R. Mohd Rasdi, M. Ismail, and B. Abu Samah, "Fostering knowledge sharing behaviour among public sector managers: a proposed model for the Malaysian public service," Journal of knowledge management, vol. 19, no. 1, pp. 121–140, 2015. https://doi.org/10.1108/JKM-11-2014-0449

[29] H. J. Mitchell, "Technology and knowledge management: Is technology just an enabler or does it also add value?," in *Knowledge management: Current issues and challenges*, IGI Global, 2003, pp. 66–78.

[30] S. Song, "An internet knowledge sharing system," Journal of Computer Information Systems, vol. 42, no. 3, pp. 25–30, 2002. https://doi.org/10.1080/08874417.2002.11647499

[31] *I. Qureshi, C. Sutter, and B. Bhatt, "The transformative power of knowledge sharing in settings of poverty and social inequality," Organization Studies, vol. 39, no. 11, pp. 1575–1599, 2018. https://doi.org/10.1177/0170840617727777*

[32] J. S. Holste and D. Fields, "Trust and tacit knowledge sharing and use," Journal of knowledge management, 2010. https://doi.org/10.1108/13673271011015615

[33] K. Niu, "Organizational trust and knowledge obtaining in industrial clusters," Journal of Knowledge Management, vol. 14, no. 1, pp. 141–155, 2010. https://doi.org/10.1108/13673271011015624

[34] R. M. Al-Adaileh and M. S. Al-Atawi, "Organizational culture impact on knowledge exchange: Saudi Telecom context," Journal of knowledge Management, vol. 15, no. 2, pp. 212–230, 2011. https://doi.org/10.1108/13673271111119664

[35] D. Nakano, J. Muniz Jr, and E. Dias Batista Jr, "Engaging environments: tacit knowledge sharing on the shop floor," Journal of Knowledge Management, vol. 17, no. 2, pp. 290–306, 2013. https://doi.org/10.1108/13673271311315222

[36] K. Blomkvist, "Knowledge management in MNCs: the importance of subsidiary transfer performance," Journal of Knowledge Management, vol. 16, no. 6, pp.904–918,2012. https://doi.org/10.1108/13673271211276182

[37] D. McNichols, "Optimal knowledge transfer methods: a Generation X perspective," Journal of knowledge management, 2010. https://doi.org/10.1108/13673271011015543

[38] W. Lam, "Barriers to e-government integration," Journal of Enterprise Information Management, 2005. https://doi.org/10.1108/17410390510623981

[39] S. Panahi, J. Watson, and H. Partridge, "Towards tacit knowledge sharing over social web tools," Journal of knowledge management, 2013. https://doi.org/10.1108/JKM-11-2012-0364

[40] M. Ranjbarfard, M. Aghdasi, P. López-Sáez, and J. E. N. López, "The barriers of knowledge generation, storage, distribution and application that impede learning in gas and petroleum companies," Journal of Knowledge Management, 2014. https://doi.org/10.1108/JKM-08-2013-0324

[41] D. Rathi, L. M. Given, and E. Forcier, "Interorganisational partnerships and knowledge sharing: the perspective of non-profit organisations (NPOs)," Journal of Knowledge Management, vol. 18, no. 5, pp. 867–885, 2014. https://doi.org/10.1108/JKM-06-2014-0256

[42] T.-M. Yang and Y.-J. Wu, "Exploring the determinants of cross-boundary information sharing in the public sector: An e-Government case study in Taiwan," J Inf Sci, vol. 40, no. 5, pp. 649–668, 2014. https://doi.org/10.1177/0165551514538742

[43] R. Zhao and B. Chen, "Study on enterprise knowledge sharing in ESN perspective: a Chinese case study," Journal of Knowledge Management, vol. 17, no. 3, pp. 416–434, 2013. https://doi.org/10.1108/JKM-12-2012-0375

[44] V. Gururajan and D. Fink, "Attitudes towards knowledge transfer in an environment to perform," Journal of knowledge Management, vol. 14, no. 6, pp. 828–840,2010. https://doi.org/10.1108/13673271011084880

[45] S. M. Mutula and J. Mostert, "Challenges and opportunities of e-government in South Africa," The electronic library, 2010. https://doi.org/10.1108/02640471011023360

[46] A. M. A. Qureshi and N. Evans, "Deterrents to knowledge-sharing in the pharmaceutical industry: a case study," Journal of Knowledge Management, 2015. https://doi.org/10.1108/JKM-09-2014-0391

[47] M. Ajmal, P. Helo, and T. Kekäle, "Critical factors for knowledge management in project business," Journal of knowledge management, vol. 14, no. 1, pp. 156–168,2010. https://doi.org/10.1108/13673271011015633

[48] A. Titi Amayah, "Determinants of knowledge sharing in a public sector organization," Journal of knowledge management, vol. 17, no. 3, pp. 454–471, 2013. https://doi.org/10.1108/JKM-11-2012-0369

[49] J. P. Meyer and L. Herscovitch, "Commitment in the workplace: Toward a general model," Human resource management review, vol. 11, no. 3, pp. 299–326, 2001. https://doi.org/10.1016/S1053-4822(00)00053-X

[50] A. Newman and A. Z. Sheikh, "Organizational commitment in Chinese small-and medium-sized enterprises: the role of extrinsic, intrinsic and social rewards," The International Journal of Human Resource Management, vol. 23, no. 2, pp. 349–367, 2012. https://doi.org/10.1080/09585192.2011.561229

[51] S. SamGnanakkan, "Mediating role of organizational commitment on HR practices and turnover intention among ICT professionals," *Journal of management research*, vol. 10, no. 1, pp. 39–61, 2010.

[52] M.-Y. Cheng, J. S.-Y. Ho, and P. M. Lau, "Knowledge sharing in academic institutions: A study of Multimedia University Malaysia.," *Electronic Journal of knowledge management*, vol. 7, no. 3, 2009.

[53] M.-C. Huang, Y.-P. Chiu, and T.-C. Lu, "Knowledge governance mechanisms and repatriate's knowledge sharing: the mediating roles of motivation and opportunity," Journal of knowledge management, vol. 17, no. 5, pp. 677–694, 2013. https://doi.org/10.1108/JKM-01-2013-0048

[54] nsf.gov, "NSF - National Science Foundation," 2023. https://www.nsf.gov/ (accessed Mar. 06, 2023).

[55] M. & C. M. Kampmann, "Knowledge Sharing Alliance: Facilitating Dialogue for Universal Development," 2014, Accessed: Mar. 07, 2023.

[Online]. Available: www.oecd.org/ knowledge-sharing-alliance

[56] "Forbes," 2023, Accessed: Mar. 08, 2023. [Online]. Available: https://www.forbes.com/places/saudi-arabia/?sh=a6fedf84e5c1

[57] N. Al Mudawi, "ACCE-GOV: a new theoretical framework for cloud computing adoption for e-government system in developing countries (Saudi Arabia perspective)," University of Sussex, 2021.

[58] S. T. H. Dulayami and L. Robinson, "The individual and the collective: Factors affecting knowledge sharing in Saudi Arabian companies," Journal of Documentation, vol. 71, no. 1, pp. 198–209, 2015. https://doi.org/10.1108/JD-09-2014-0121

[59] "Vision 2030," 2023, Accessed: Mar. 08, 2023. [Online]. Available: https://www.vision2030.gov.sa/

[60] F. Ghabban, A. Selamat, and R. Ibrahim, "New model for encouraging academic staff in Saudi universities to use IT for knowledge sharing to improve scholarly publication performance," Technol Soc, vol. 55, pp. 92–99,2018. https://doi.org/10.1016/j.techsoc.2018.07.001

[61] "Tabadul," 2023, Accessed: Mar. 07, 2023. [Online]. Available: https://www.elm.sa/en/about/Pages/Tabadul.aspx

[62] S. Q. A.-K. Al-Maliki, "Information and communication technology (ICT) investment in the Kingdom of Saudi Arabia: Assessing strengths and weaknesses," Journal of Organizational Knowledge Management, vol. 2013, p. 1, 2013. https://doi.org/10.5171/2013.450838

[63] F. Yassin, J. Salim, and N. Sahari, "The influence of organizational factors on knowledge sharing using ICT among teachers," Procedia technology, vol. 11, pp.272–280,2013. https://doi.org/10.1016/j.protcy.2013.12.191

[64] M. Al-Khalifa, S. Khatoon, A. Mahmood, and I. Fatima, "Factors influencing patients' attitudes to exchange electronic health information in Saudi Arabia: an exploratory study," *International Journal of Advanced Computer Science and Applications*, vol. 7, no. 8, 2016.

[65] D. Chandran and A. M. Alammari, "Influence of culture on knowledge sharing attitude among academic staff in eLearning virtual communities in saudi Arabia," Information Systems Frontiers, vol. 23, pp. 1563–1572, 2021. https://doi.org/10.1007/s10796-020-10048-x

[66] R. Fullwood and J. Rowley, "An investigation of factors affecting knowledge sharing amongst UK academics," Journal of Knowledge Management, 2017. https://doi.org/10.1108/JKM-07-2016-0274

[67] F. M. Alsaadi, "Knowledge Sharing Among Academics in Higher Education Institutions in Saudi Arabia," 2018.

[68] A. N. Tashkandi and I. M. Al-Jabri, "Cloud computing adoption by higher education institutions in Saudi Arabia: an exploratory study," Cluster Comput, vol. 18, pp. 1527–1537, 2015. https://doi.org/10.1007/s10586-015-0490-4

[69] A. Ali, "Cloud computing adoption at higher educational institutions in the KSA for Sustainable Development," *International Journal of Advanced Computer Science and Applications*, vol. 11, no. 3, 2020.

[70] Quentin L. Burrell, ", Isle of Man International Business School, Isle of Man. Definition 7 on p. 481 of Zins, C. (2007). Conceptual Approaches for Defining Data, Information, and Knowledge. JASIST, 58, 479-49," p. 7, 2023.

[71] Raya Fidel, ", University of Washington, Seattle, WA. Definition 17 on page 483 of Zins, C. (2007). Conceptual Approaches for Defining Data, Information, and Knowledge. JASIST, 58, 479-493.," 2023.

[72] William Hersh, ", Oregon Health Science University, Portland, OR. Definition 24 on page 484 of Zins, C. (2007). Conceptual Approaches for Defining Data, Information, and Knowledge. JASIST, 58, 479-493," 2023.

[73] Charles Oppenheim, ", Loughborough University, Leicestershire, UK. Definition 32 on page 485 of Zins, C. (2007). Conceptual Approaches for Defining Data, Information, and Knowledge. JASIST, 58, 479-493.," 2023.

[74] Richard Smiraglia, ", Long Island University, Brookville, NY. Definition 38 on page 486 of Zins, C. (2007). Conceptual Approaches for Defining Data, Information, and Knowledge. JASIST, 58, 479-493," 2023.

[75] Anna da Soledade Vieira, ", Federal University of Minas Gerais, Belo Horizonte, Minas Gerais, Brazil. Definition 42 on page 486 of Zins, C. (2007). Conceptual Approaches for Defining Data, Information, and Knowledge. JASIST, 58, 479-493.," 2023.

[76] Donald Hawkins, ", Information Today, Medford, NJ. Definition 21 on p. 483 of Zins, C. (2007). Conceptual Approaches for Defining Data, Information, and Knowledge. JASIST, 58, 479-493.," 2023.

[77] Donald Kraft, ", Louisiana State University, Baton Rouge, LA. Definition 25 on p. 484 of Zins, C. (2007). Conceptual Approaches for Defining Data, Information, and Knowledge. JASIST, 58, 479-493.," 2023.

[78] Aldo de Albuquerque Barreto, ", Brazilian Institute for Information in Science and Technology, Brazil. Definition 3 on p. 480 of Zins, C. (2007). Conceptual Approaches for Defining Data, Information, and Knowledge. JASIST, 58, 479-493.," 2023.

Position Papers of the 18th Conference on Computer Science and Intelligence Systems pp. 17–26
DOI: 10.15439/2023F4109
ISSN 2300-5963 ACSIS, Vol. 36

ACC-PH: a Comprehensive Framework for Adopting Cloud Computing in Private Hospitals

Fayez Alshahrani
0009-0001-7389-0548
Department of Information Systems, Najran University, Najran, Saudi Arabia
Department of Informatics, University of Sussex, Brighton, United Kingdom
Email: fa461@sussex.ac.uk

Natalia Beloff
0000-0002-8872-7786
Department of Informatics, University of Sussex, Brighton, United Kingdom
Email: n.beloff@sussex.ac.uk

Martin White
0000-0001-8686-2274
Department of Informatics, University of Sussex, Brighton, United Kingdom
Email: m.white@sussex.ac.uk

***Abstract*—The healthcare sector is of paramount importance as it provides necessary medical services to sustain human lives. In the private healthcare sector, organisations place equal emphasis on profits as on providing essential medical services. Thus, to offer optimal health aids at low cost, private healthcare organisations try to acquire the best technologies available. Cloud computing offers a solution to cutting business expenses while boosting productivity because it supplies computing services through third parties more cost-effectively. Nonetheless, recent studies have shown that adopting cloud computing services in private healthcare facilities in Saudi Arabia is behind when compared to other sectors. This study presents an optimal data collection and framework validation methodology, combining qualitative and quantitative approaches to examine proposed factors influencing Adopting Cloud Computing in Private Hospitals (ACC-PH) in Saudi Arabia. Accordingly, this research is expected to enhance the implementation of cloud computing in Saudi private hospitals.**

***Index Terms*—cloud computing, technology adoption models, Saudi private healthcare sector.**

I. Introduction

THE healthcare sector is an essential pillar of all civic assemblies worldwide. They supply imperative conditions for human existence on Earth. Accordingly, they seek to keep pace with all developments to render exemplary medical services. This incorporates a shift from traditional health systems management to modern electronic systems. Thus, in light of their continuous endeavour toward development, hospitals have tried moving toward implementing Information and Communication Technologies. That movement is called Health Systems.

E-Health systems have provided numerous advantages for healthcare organisations and their stakeholders. They have facilitated sharing of data for doctors, patients, and other health practitioners [1]. Implementing E-Health systems has also raised healthcare quality and safety [2]. In addition, several E-Health programmes have been used to enhance education and positively change the behaviour of many patients and health practitioners [2]. However, despite the dramatic paradigm shift in the medical care field caused by the utilisation of E-Health systems, these systems have undergone challenges vastly.

The higher cost and shortage of technical experts primarily hinder the implementation of E-Health systems in hospitals. These obstacles further complicate the situation, particularly for the private healthcare sector, which is more concerned with profitability principles. Technologies used in E-Health systems need periodic maintenance, updates, and technical support, which increases the financial burden on healthcare organisations [2]. Also, healthcare organisations face a significant barrier to the availability of skilful technicians [3]. As technicians' presence operates E-Health systems, their absence is a hurdle. Consequently, hospitals ought to employ contemporary technologies to resolve such issues.

Cloud computing is the desired innovation that can help hospitals overcome their E-health problems. Cloud computing removes upfront capital investments in technical infrastructure and maintenance from hospitals' shoulders to be cloud providers' responsibility [4]. Cloud adoption benefits further reduce the need for technical experts, whose availability is essentially a dilemma facing hospitals. This will also indirectly contribute to a reduction in costs associated with recruiting technicians. And given the profitability model of private hospitals, using cloud technology in E-Health programmes is a compelling solution. It would therefore be surprising not to embrace cloud computing in this vital sector in countries that want inclusive development and economic prosperity.

In Saudi Arabia, which seeks a digital transformation to achieve an extraordinary economic renaissance according to a new vision, the private healthcare sector remains technically underdeveloped. Some private hospitals in Saudi Arabia still generate health records only in paper versions [5]. This makes it seem like certain private hospitals in Saudi Arabia will never make the leap to more cutting-edge technologies like

©2023, PTI

Topical area: Information Technology for Business and Society

cloud computing. Previous research has revealed that the use of cloud technology in the Saudi medical care sector, including the private industry, is the lowest in the nation [5]. Given the critical role of these hospitals in Saudi Arabia, their reluctance to adopt suitable technologies poses a danger.

The latest recent data from Saudi Arabia's Ministry of Health show that private hospitals served more than 43 per cent of patients in the country in 2021 [6]. Hence, it is of the utmost importance that these types of hospitals continue to exist and grow to serve a large portion of the Saudi population with necessary medical care. This necessary sustainability and development drive Saudi private hospitals to catch up with technical development by adopting appropriate modern technologies such as cloud computing.

Therefore, this study aims to help Saudi private hospitals in adopting cloud computing by establishing a framework to comprehend the most effective aspects influencing the adoption decision. The remainder of this research will be constructed, beginning with a literature review to identify elements that have proven influence in a similar context. The construction of the comprehensive framework presented in this study comes next. Finally, the study will be ended by determining the most appropriate methodologies for data collection and validating the proposed model.

II. LITERATURE REVIEW

Prior studies on cloud technology adoption topics in the Saudi healthcare industry are investigated in this section. The discussion includes the studies of [7]–[10]. This analysis seeks to discover determinants that have proven to affect cloud technology implementation in Saudi hospitals. Consequently, these factors will serve as primary pillars of a comprehensive framework designed to assist private hospitals in Saudi with cloud computing's successful utilisation.

All the papers analysed in our research focused on the technological context of analysing potential implications from this perspective. Prior research in the Saudi healthcare industry indicated positive effects of relative advantage, compatibility, security, and reliability as technological factors.

The relative advantage factor, which means the extent to which adopting the cloud raises the efficiency of other current technologies in institutions, has emerged as an influential factor in the studies of [7], [10]. In contrast, concerns of healthcare institutions in Saudi Arabia towards security issues with the cloud have shown its impact in the other two studies [8], [9]. In addition, [7]–[9] argued that adopting cloud technology in Saudi hospitals is positively affected by cloud compatibility with healthcare organisations' current policies and principles. Moreover, the reliability of providing cloud services with no interruptions positively impacts the adoption in Saudi hospitals as solely approved by [8].

Although there are apparent conflicts in the findings, as shown in Table I., the literature showed the importance of some technological factors. However, the technological context was not the only context influencing cloud computing adoption.

TABLE I.
THE CLOUD TECHNOLOGY DEPLOYMENT IN SAUDI'S HEALTHCARE SECTOR: TESTED VARIABLES

	Reference [7]	**Reference [8]**	**Reference [9]**	**Reference [10]**
Relative Advantage	√	Ø	×	√
Technology Readiness	Ø	Ø	Ø	×
Compatibility	√	√	√	×
Complexity	×	Ø	×	×
Regulations and Rules	×	Ø	×	√
Competitive Pressure	×	Ø	√	Ø
External Expertise	Ø	Ø	×	×
Costs Analysis	Ø	Ø	Ø	√
Top Management Support	√	Ø	√	×
Attitude toward Change	Ø	√	×	√
Internal Expertise	Ø	Ø	Ø	×
Prior Experience	√	Ø	×	×
Security	Ø	√	√	Ø
Organisational Readiness	×	Ø	√	Ø
Data Control	Ø	×	Ø	Ø
Data Privacy	Ø	×	Ø	Ø
Reliability	Ø	√	Ø	Ø

* × = negative impact, √ = positive impact, Ø = not investigated.

Several crucial variables originated from the organisation's scope and positively influenced cloud technology deployment in Saudi medical centres. Ayadi's [9] research indicated that organisational readiness with required human, technological, and financial resources positively affected cloud adoption in Saudi hospitals. Human resources, in particular, have played a prominent role in shaping influences of adopting from the organisational context.

Human resources impact was proved in [7], [9], which revealed a positive effect of involvement and support from senior executives (top management support) on the decision to embrace the cloud. Their prior technological experience is another significant factor, as [7] claimed. In addition, [8], [10] emphasised that the desire, feeling, and orientation of workers, particularly Information Technology (IT) department employees, towards the new technology was a major factor in the cloud adoption decision. However, this factor and others were studied in contexts other than organisational context.

The attitude towards change was not investigated by [8] as an organisational but technological factor. However, we believe this factor relies on human elements that are an integral part of an organisation, meaning it must be considered in that setting. The same point of view applies to another variable that has strongly influenced cloud migration in Saudi hospitals and was addressed from a business perspective. This factor is cost analysis which refers to the extent that analysing returned benefits and expense cost of adopting cloud computing can affect the adoption decision [10]. We argue that cost analysis is carried out by relevant employees within healthcare organisations, which places it in the organisational context. While we place cost analysis as a determining factor within organisational influencers, prior research has identified other influencers originating from external contexts.

The literature revealed some influencing factors on the cloud adoption decision came from the surrounding environment of Saudi healthcare organisations. According to [9], Saudi hospitals' use of the cloud is largely influenced by industry competition as they are inspired by competitors' use of cloud computing and its unique benefits. Another environmental factor is the impact of government rules and regulations which was proven in a study conducted by [10]. However, the impact of environmental context was rejected or neglected in the other studies.

Almubarak's [7] study did not show the influence of any environmental factors. And a study by [8] did not investigate the environmental context at all. However, this cannot be an argument not to investigate this crucial context. Specifically, this study is situated within the context of a developing country, necessitating the examination of critical environmental factors, notably cloud providers and Internet connection, which have not been investigated in prior studies.

In short, this literature review contributes to highlighting the most significant influencing factors in the same context of our research. It also highlighted significant gaps represented in the lack of research targeting the Saudi healthcare sector and its total absence in the private healthcare sector. In addition, prior studies have been unable to analyse the impact of some factors that cannot be ignored, particularly in the setting of a developing nation like Saudi Arabia. Therefore, to fill these gaps, developing a novel comprehensive framework for cloud deployment in Saudi private hospitals becomes an essential necessity.

III. Framework

This section discusses the development of a comprehensive framework for Adopting Cloud Computing in Private Hospitals (ACC-PH) in Saudi Arabia. The framework is going to be constructed by merging two of the most significant theories in deploying new technology: the theory of Technology, Organisation, and Environment (TOE) and the model of Diffusion of Innovation (DOI).

A. The Theory of Technology, Organisation, and Environment

Tornatzky [11] established the model of the TOE to determine constraints and opportunities that influence adopting an innovation within an organisation. It considers technical and non-technical issues, like the surrounding environment and internal structure. The TOE framework examines how an organisation embraces and executes a new technology and how the technology, organisation, and environmental contexts can influence that adoption [10]. The significant advantage of TOE is that it provides researchers with an open land in which they can categorise features based on each circumstance within a wide sphere [7]. However, this theory suffers some glaring flaws that cannot be ignored. Researchers have contended that TOE does not account for all variables in every context; for example, cloud computing requires multiple conceptual approaches to articulate a fuller insight into the adoption choice [2].

The TOE theory alone is insufficient to determine all factors impacting the choice to uptake cloud technology in the healthcare industry; hence, additional approaches must supplement it. Many studies used the framework of TOE along with the model of Diffusion of Innovation (DOI) to handle the technological and operational challenges of embracing cloud technology involving the medical care sector [12].

B. The Model of Diffusion of Innovation

The DOI model is the most often used in conjunction with the TOE framework because they complement one another and are used together in the Saudi healthcare context [7]. Roger [13] developed the DOI model to evaluate variables that influence innovations' deployment. According to Roger's theory, adopting technology involves numerous steps, from initial awareness of innovation to accepting or refusing, executing, and affirming the decision. The DOI theory states that each innovation has a set of features that affect its prevalence, and these features are: "relative advantage, compatibility, complexity, trialability, and observability" [14].

Relative advantage examines the impact of an innovation on the system, compatibility assesses how well the innovation aligns with current systems, complexity determines the usability of the technology, trialability analyses the accessibility of the innovation, and observability reflects the innovation's visibility level [15]. According to previous studies, relative advantage and compatibility have proven to affect the selection process to uptake cloud technology. They will therefore be taken to form the ACC-PH framework factors proposed in this research.

C. The Comprehensive Framework for Adopting Cloud Computing in Private Hospitals

Most of the ACC-PH framework components presented in Fig. 1 will be derived from the literature as they proved their positive impact in the same context. In addition, other factors will be added to bridge the lacunae in literature and develop an essential comprehensive framework. The ACC-PH framework will investigate novel variables not yet explored in the

research literature within our research context. These factors are ***cloud providers*** and ***Internet connection.***

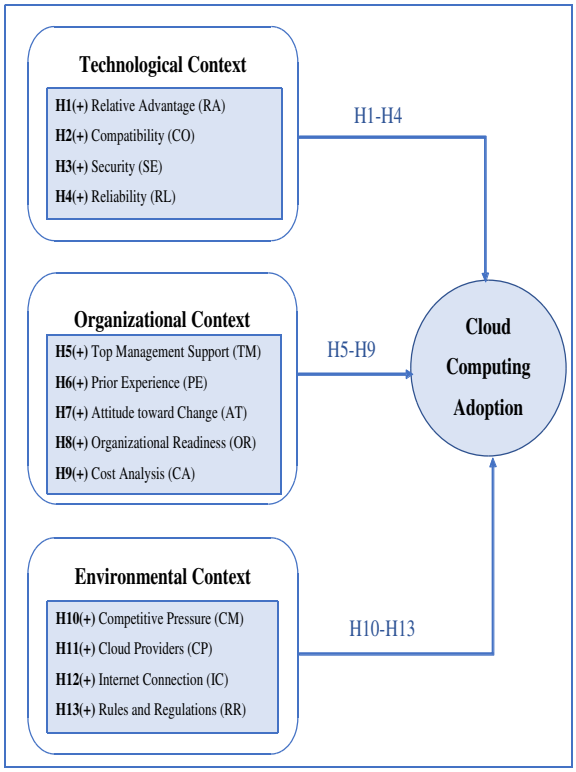


Fig. 1 The Comprehensive Framework for Adopting Cloud Computing in Private Hospitals (ACC-PH)

Cloud providers and ***Internet connection*** factors considerably increase the likelihood of utilising cloud technology in the developing world. However, they have not been sufficiently investigated in Saudi Arabia's medical care sector. Hence it is essential to analyse these aspects in Saudi Arabia, which is a country characterised as a developing nation, where cloud service providers are scarce [16] and Internet connectivity is limited [17]. Therefore, adding these factors to the ones that have demonstrated a positive influence in earlier research will undoubtedly contribute to developing a comprehensive framework to facilitate embracing cloud technology in the Saudi Private medical care industry. Importantly, each variable will be examined in its proper context and based on the established TOE theory, which includes technological, organisational, and environmental aspects. Thus, all ACC-PH factors will be analysed and hypothesised within these three contexts.

1) The Technological Context

a. Relative Advantage

The Relative Advantage (RA) factor describes the extent to which a hospital can acquire additional advantages by utilising cloud computing [9]. Organisations are captivated by the cloud because of its decentralised structure, instantaneous provisioning of computing resources, and expansive storage capabilities [18]. In addition, cloud systems are cost-effective, so businesses are more likely to invest in cutting-edge technology that contributes directly to their bottom line [19]. Furthermore, healthcare decision-makers widely believe that implementing cloud computing can improve medical services by increasing responsiveness and reducing technological infrastructure failures [7].

Although the deployment of cloud computing in healthcare facilities may provide significant benefits, it is crucial to acknowledge the presence of some risks associated with its adoption. The storage of confidential medical information on remote computer systems raises concerns about privacy and security [20]. The lock-in effect poses additional hazards that result in cloud users becoming dependent on the services offered by the cloud service provider, so limiting customers' flexibility in their choices [21]. In addition, the adoption of cloud computing has the potential to amplify the risk of data loss and exacerbate system downtime when utilising an inadequate cloud provider [22]. Moreover, cloud adoption could cause compliance problems if the cloud provider violates data protection and healthcare sector requirements [23], [24].

However, healthcare organisations can mitigate these risks in order to optimise the considerable advantages offered by cloud computing. To protect hospital operations and patient data, hospitals should conduct risk assessments, evaluate multiple cloud providers, review service level agreements, encrypt data, restrict access, and create contingency plans [25], [26]. These proactive measures can bolster their security posture and mitigate potential risks associated with cloud adoption.

Therefore, it is of the utmost importance to investigate the relative advantage factor as a positive element in deploying cloud technology in private medical facilities in Saudi Arabia. Hence the following hypothesis is proposed.

H1: Recognising the relative advantage of cloud technology enhances the likelihood of adopting cloud computing in Saudi private hospitals.

b. Compatibility

Compatibility (CM) refers to evaluating how well the to-be-adopted technology aligns with the enterprise's current values, historical experience, and operational requirements [13]. Some perceive compatibility as a crucial factor for encouraging the implementation of technologies that are deemed to align with present organisations' strategies, principles, and employee abilities [27]. Precisely, compatibility in the view of cloud adoption confronts more unique requirements, which include ensuring the simplicity of exchanging data between traditional and cloud systems (integration) and the adjustment of services (customisation) [28].

However, many believe that compatibility issues hinder widespread technological adoption, including cloud computing [29], [30]. Yet, this argument could be obscured by numerous studies that have validated the significance of the compatibility factor by demonstrating its impact on cloud acceptance [27], [31]. Therefore, to experience the advantages of cloud adoption, medical institutions must realise the necessity of compatibility and implement necessary up-

grades to the current system. Based on that, we posit compatibility as a key influencing variable in deploying cloud computing in Saudi private hospitals. The next hypothesis is, therefore, proposed.

H2: Higher compatibility enhances the likelihood of adopting cloud computing in Saudi private hospitals.

c. Security

Security (SE) refers to how well a cloud provider safeguards and protects the data of both users and organisations [32]. Security cannot be overlooked when implementing cloud technology in the healthcare industry, which varies significantly from other sectors due to its primary security and privacy concerns that can sometimes be life-threatening [12]. In current E-Health systems, ensuring the safety and privacy of healthcare data remains crucial even without migrating to an Internet-based technology such as the cloud [33]. Adopting cloud computing in E-Health systems raises risks associated with trust and confidentiality of data sharing [34]. Integrity and availability are other security concerns related to migrating E-Health systems to the cloud [20].

The preservation of confidentiality, integrity, and availability has significant importance in ensuring robust cloud security within E-Health systems. Ensuring the security of confidential health-related information, preserving data accuracy, and guaranteeing uninterrupted access to data and applications are fundamental aspects [20], [34]. Protecting these vital components in the cloud calls for outstanding encryption, access restrictions, data integrity mechanisms, and a solid infrastructure [20].

Therefore, security must maintain high standards to encourage hospitals to benefit from cloud computing and its numerous benefits. Cloud computing can enable a highly secure software environment [35]. However, healthcare institutions must ensure that cloud providers implement proper security measures [10].

Providers of cloud services have an important part to play in helping the healthcare industry conform to regulations and safeguard sensitive patient data. To comply, they must establish strong access restrictions, encryption techniques, perform frequent security audits, and have disaster recovery plans in place [23]. Accurate documentation and independent verification boost healthcare organisations' confidence and encourage wider use of cloud computing [36].

However, the price of moving to the cloud may rise if strict security measures are implemented [37]. There are a number of factors that contribute to this, including strong access restrictions, encryption, security inspections, external certifications, disaster recovery, and data redundancy.

Therefore, healthcare organisations should collaborate with cloud service providers to strike a good balance between security levels and economic efficiency [37]. Various levels of security tiers may be accessible, and allocating resources towards comprehensive security measures might result in enduring advantages, mitigating the likelihood of data breaches and minimising regulatory penalties [38]. Ultimately, this investment is necessary to maintain the confidentiality, availability, and integrity of sensitive healthcare data in the cloud.

Thus, we assert it is pertinent that ensuring high-level security standards will increase acceptance of cloud technology deployment in the Saudi private medical care industry. The next hypothesis is, thus, proposed.

H3: Higher security levels enhance the likelihood of adopting cloud computing in Saudi private hospitals.

d. Reliability

Reliability (RL) refers to a system's capacity to accomplish expected tasks, including delivering high-quality services at high speed while lowering failure rates and offering data recovery in case of such failures [39]. In healthcare delivery, system reliability takes on a greater significance. Access to accurate and trustworthy patient data is crucial for healthcare providers [40]. Thus delays in responding to urgent instances that could endanger a patient's life highlight the increasing importance of computing environments with scalable capabilities and reliability [41].

Therefore, cloud computing's capacity to maintain a highly dependable computing environment has the potential to improve healthcare delivery significantly [42]. Since data are kept and accessed from numerous distinct places, the system will not be down if one server crashes, which is the essential characteristic of cloud computing [41].

Accordingly, considerable research has demonstrated that the reliability variable is crucial and positively influences cloud computing deployment, particularly in the Saudi medical care industry [8], [39]. Hence the next hypothesis is proposed.

H4: Higher reliability enhances the likelihood of adopting cloud computing in Saudi private hospitals.

2) The Organisational Context

a. Top Management Support

Top Management Support (TM) is meant by the degree to which an organisation's leaders are behind the initiative to embrace innovation and give it the resources it needs to succeed [7]. Top executives are taking their important role from being the ones who are in charge of making critical decisions [43]. A recent study identified senior management's acceptance and understanding of new technology as a critical factor influencing adoption intentions [44]. It is perceived that if upper-level management realises the value of the technology, they will encourage their employees to utilise it. If they do not, they will act as a roadblock [45]. In addition, when leaders can understand what is needed to implement cloud computing and how to accomplish it, they can then provide human resources and equipment essential for cloud adoption [46].

Therefore, the decision to implement new technologies within organisations is heavily influenced by the support of top management [47]. Numerous studies have indicated that the level of support from high management influences cloud computing embracing [48], [49], including Saudi healthcare

organisations [7], [10]. Therefore, the next hypothesis can be proposed.

H5: Top management support enhances the likelihood of adopting cloud computing in Saudi private hospitals.

b. Prior Experiences

Prior Experiences (PE) allude to the degree to which the impact of past technical skills influences the choice to adopt new technology [7]. The lack of technological expertise among decision-makers exacerbates the resistance to changes towards up-taking technology like cloud computing [19]. Some studies suggest that the absence of technical expertise among top executives regarding cloud computing impedes adoption [43]. In contrast, top managers' familiarity with the technical aspects of cloud computing can help them see the potential advantages of making the switch [7].

In the Saudi medical care sector, the influence of top managers' prior experience has positively impacted cloud computing deployment [7]. Thus, the next hypothesis can be proposed.

H6: Top managers' sufficient prior technical experience enhances the likelihood of adopting cloud computing in Saudi private hospitals.

c. Attitude towards Change

Attitude towards Change (AT) alludes to the degree to which favourable or unfavourable emotions, cognitive aspects, behavioural control, and beliefs that employees hold about new technology can affect embracing innovation [8], [44]. Employees are firms' primary pillars of introducing new technology [47]. This is certainly relevant within information technology departments. They are the ones who are accountable for the majority of the adoption process. If they have a negative attitude, they will not put forth their best effort in training, learning, and engaging in the adoption [47].

Some ascribe resistance to adopting new technology to the extensive training required for employees to grasp the skills necessary to work with these systems [8]. Others, however, believe that this training is crucial for preparing workers within institutions to adopt a more optimistic stance towards emerging innovations, hence facilitating the adoption process [48].

Therefore, the attitude towards change can positively influence new technology deployment. And this has been proven in past research that examined the deployment of cloud technology in Saudi private medical care institutions [8], [10]. Hence, the next hypothesis is proposed.

H7: Positive employees' attitudes towards change enhance the likelihood of adopting cloud computing in Saudi private hospitals.

d. Organisational readiness

Organisational Readiness (OR) refers to how the availability of people, financial, and technological resources and an organisation's ability to absorb and apply new information and knowledge can influence the acceptance of new technology [7], [28], [50]. Transforming non-technological institutions, like healthcare institutions, from traditional E-Health systems to modern cloud systems requires readiness to absorb knowledge and information associated with these technologies. In addition, access to sufficient human, financial, and technological resources is crucial for modernising healthcare organisations with such a transition, as it directly influences the ease of adopting and implementing new technology within an institution [51].

Research has examined how much organisational readiness, including human, financial, and technological resources, can affect cloud computing adoption. Specifically, this factor has been investigated at the level of healthcare institutions in Saudi Arabia and has demonstrated its influence on cloud technology development [7]. The next hypothesis is, therefore, proposed.

H8: Organisational readiness enhance the likelihood of adopting cloud computing in Saudi private hospitals.

e. Cost Analysis

The Cost Analysis (CA) factor refers to the extent to which a comprehensive organisation-wide evaluation of benefits versus anticipated costs of deploying new cloud technology can impact the adoption decision [2]. The perceived usefulness of a technological innovation hinges on the balance between its advantages and expenditures, causing businesses constant concern over the price tag of adopting cutting-edge innovations [52]. It is no secret that the staggering expense of upgrading to modern technologies is a major obstacle for organisations when it comes to IT deployment and acceptance [4].

The deployment costs of cloud computing adoption can include capital expenditure (CAPEX), which includes one-time costs and operating expenditure (OPEX), which includes recurring costs [53]. The capital expenses of deploying cloud computing may include the cost of building up network and Internet connections, while the operating costs pertain to data transfers between customers and providers of the cloud [53]. Based on [2], for organisations seeking the utilisation of cloud technology, "The cost should be analysed in both capital expenditure (CAPEX) and operational expenditure (OPEX)". Moreover, it is essential for the cost analysis to consider all relevant elements that impact the decision-making process for potential cloud clients.

The phenomenon of customer lock-in can provide additional complexity to the decision-making process of adopting a certain product or service. This is particularly relevant in cases when certain features are only provided by a specific vendor and are not accessible through open-source alternatives [54]. This can cause consumers to become reliant on a single service provider, restricting their flexibility and options.

Furthermore, the possible hazards associated with the removal of key services from the market raise substantial apprehensions for organisations, particularly in vital sectors such as healthcare [55]. Within the context of a hospital's daily operating processes, the absence of specific essential services might result in interruptions and difficulties in identifying appropriate substitutes.

On the other hand, organisations are increasingly turning to cloud computing as an on-demand IT service model to cut costs and increase productivity [12]. Cloud computing eliminates the need for initial and ongoing investments in an organisation's information technology infrastructure, lowering overall costs [8]. Cloud computing technology helps organisations save money on start-up expenses by eliminating the need to purchase and set up expensive hardware to power computing services [56].

Therefore, weighing the advantages and drawbacks of using cloud technology in terms of cost analysis as a potential influencer to the decision to adopt the cloud is important. Numerous past research has shown the positive impact of analysing cost on cloud adoption [10], [52]. Therefore, the next hypothesis is proposed.

H9: Efficient cost analysis enhances the likelihood of adopting cloud computing in Saudi private hospitals.

3) The Environmental Context

a. Competitive Pressure

Competitive Pressure (CP) alludes to the influence that rival enterprises within an industry have on whether or not a certain institute adopts a given invention [7], [57]. Competitors using cutting-edge technology exert pressure on non-adopters in the same industry by enhancing work productivity and delivering superior services [58]. Also, competitors exert influence on executives to deploy innovative technologies like cloud computing to meet market demands and enhance operational effectiveness [59].

Prior studies have found that the competitive pressure variable effect on the deployment of cloud technology ranged from being natural to having a negative or positive impact. Gutierrez' [60] study concluded that the competitive pressure variable did not impose any influence on Cloud technology utilisation, while [7] demonstrated the neutrality of the factor impact. However, the vast majority of research has proven the significant positive influence of the competitive pressure factor in adopting cloud systems [9], [10], [61]. The next hypothesis is, thus, proposed.

H10: Competitive pressure enhance the likelihood of adopting cloud computing in Saudi private hospitals.

b. Cloud Providers

The Cloud Providers (CP) factor refers to how the availability and location of providers impact organisations' decisions regarding the adoption of cloud computing. Cloud providers are third-party individuals, businesses, non-profits, or government agencies that store data for another organisation [62]. Since cloud providers are responsible for most cloud functions, they also have the primary burden of ensuring the safety of stored data [63]. The fundamental cause for concern regarding cloud providers is the location where the information is stored.

Users who store sensitive information, like health information, in the cloud generally desire to be aware of the location at which their data are stored. They may prefer to choose a particular location [64]. However, many cloud users are unaware of the cloud provider's data centre location where their data are stored [65]. Data may be maintained in a country other than the one where the adopting institution is situated, contributing to increased threats to data security and privacy and complications of different laws governing data [66]. Hence, data location is a core principle of successful cloud computing deployment [67]. Therefore, we propose that the existence of a cloud provider in the same country as cloud consumers can increase the likelihood of adoption.

In the context of this research, there is a severe shortage of providers of the cloud in Saudi Arabia [16]. Therefore, adding new service providers in the country may positively affect the deployment of this innovation, particularly by institutes extremely concerned about data security and privacy, such as healthcare institutions. Thus, the next hypothesis is proposed.

H11: Cloud providers' availability within the same country enhances the likelihood of adopting cloud computing in Saudi private hospitals.

c. Internet Connection

The Internet Connection (IC) factor refers to how the availability of a capable Internet connection influences organisations' decisions to adopt cloud computing. Access to the Internet is crucial for the functionality of cloud technology, as it leverages Internet technology to deliver computer services to users [68], [69]. Cloud technology must maintain a high-speed, stable, and highly accessible connection at an affordable price [70]. If the Internet connection is inadequate, the benefits of employing cloud computing in businesses will be non-existent [17]. Thus, a fast and reliable Internet connection is essential as it is the primary link between clients and providers within the cloud, ensuring optimal speed and performance [71].

Several studies indicate that network connection is the most effective element in cloud technology deployment, particularly in developing nations [72], [73]. Specifically, in Saudi Arabia, the Internet connection was a significant worry for organisations planning to adopt cloud computing [74]. Some high-level Saudi executives complain about the poor infrastructure for Internet connections and the high cost of the Internet [17].

Therefore, before developing cloud systems, suppliers of Internet services should seek to enhance Internet architecture and ensure full functionality for the Internet backbone service at an affordable price [17]. Thus, improving Internet connectivity infrastructure by communications companies in Saudi Arabia may enhance the decision likelihood of private hospitals to deploy cloud technology. The next hypothesis is, hence, proposed.

H12: Internet connection availability and high functionality enhance the likelihood of adopting cloud computing in Saudi private hospitals.

d. Rules and Regulations

The Rules and Regulations (RR) factor means the extent to which government policies governing the utilisation of a given innovation influence institutions' adoption of this technology, particularly in data security and privacy standards [2]. In every country, the operation of any technology is regulated

by establishing explicit policies to protect its users. Regulatory policies become more urgent and essential when enterprises with a high data sensitivity and privacy level, such as healthcare institutions, implement modern technology [12]. As a result, there is a potential obstacle to technology adoption in the healthcare sector without appropriate rules and standards [2], [75].

However, in 2014, the government of Saudi appointed a special commission to create the required cloud computing rules [76]. In addition, the Ministry of Communication and Information Technology in Saudi Arabia updated the cloud computing rules and regulations in 2020 [77]. The new objective for the public and private organisations in the nation is to "accelerate the adoption of cloud computing services by directing these entities to consider cloud options when making new IT investment decisions".

Therefore, we propose that flexible rules and regulations established by the Saudi government can positively encourage cloud computing deployment. Thus, the next hypothesis is proposed.

H13: Flexible rules and regulations enhance the likelihood of adopting cloud computing in Saudi private hospitals.

IV. CONCLUSION AND FUTURE WORK

This position study developed a comprehensive framework for Adopting Cloud Computing in Private Hospitals (ACC-PH) in Saudi Arabia. The developed model is based on merging the well-known theories of TOE and DOI. The proposed variables in the model were largely derived from prior studies in the same context to produce a new framework approach. In addition, some other factors were nominated to be tested in the Saudi healthcare sector for the first time. All factors were then grouped to be explored under three primary contexts: technology, organisation, and environment. The next step in developing our framework will involve collecting and analysing data to validate the framework.

A mixed methods approach is proposed since it gives a comprehensive understanding of the studied problem by IS researchers, allowing us to collect the data necessary to validate the proposed framework [78]. Using a qualitative method to discover the most effective variables in embracing cloud technology in Saudi private hospitals: and then examining the influencing factors in this adoption by conducting a quantitative approach will enhance the accuracy of the findings. The design of questionnaires, which is the preferred method for collecting quantitative data, will be informed based on the analysis of interviews, which serves as the preferred method for collecting qualitative data.

Semi-structured interviews will be conducted to explore and validate the proposed framework. Eight to ten interviews will be in-person or online mode. The targeted sample for the interviews will be only the hospital directors, medical directors, and heads of IT departments. The questionnaire will then be structured and distributed through email and Google Forms. The questionnaire's target sample size is at least 400 participants, including administrative at Saudi private hospitals. In short, the research will combine interviews and questionnaires as the optimal methods for gathering the data necessary to evaluate the new cloud computing adoption framework (ACC-PH).

REFERENCES

[1] R. Sivan and Z. A. Zukarnain, "Security and privacy in cloud-based e-health system," *Symmetry*, vol. 13, no. 5. MDPI AG, 2021. doi: 10.3390/sym13050742.

[2] F. Alharbi, A. S. Atkins, C. Stanier, and A. Atkins, "Strategic framework for cloud computing decision-making in healthcare sector in Saudi Arabia," in *The Seventh International Conference on eHealth, Telemedicine, and Social Medicine*, 2015, pp. 138–144.

[3] S. C. Chang, M. T. Lu, T. H. Pan, and C. S. Chen, "Evaluating the e-health cloud computing systems adoption in Taiwan's healthcare industry," *Life*, vol. 11, no. 4, 2021, doi: 10.3390/life11040310.

[4] M. Singh, P. K. Gupta, and V. M. Srivastava, "Key challenges in implementing cloud computing in Indian healthcare industry," in *2017 Pattern Recognition Association of South Africa and Robotics and Mechatronics (PRASA-RobMech)*, 2017, pp. 162–167. doi: 10.1109/RoboMech.2017.8261141.

[5] S. T. Alharbi, "Users' acceptance of cloud computing in Saudi Arabia: An extension of Technology Acceptance Model," *International Journal of Cloud Applications and Computing* , vol. 2, no. 2, pp. 1–11, 2012, doi: 10.4018/ijcac.2012040101.

[6] MOH Statistical Yearbook, "Statistical Yearbook," 2021. [Online]. Available: https://www.moh.gov.sa/en/Ministry/Statistics/book/Documents/Statistical-Yearbook-2021.pdf

[7] S. S. Almubarak, "Factors influencing the adoption of cloud computing by Saudi university hospitals," *International Journal of Advanced Computer Science and Applications (IJACSA)* , vol. 8, no. 1, 2017, doi: 10.14569/ijacsa.2017.080107.

[8] M. O. Alassafi, "Success indicators for an efficient utilisation of cloud computing in healthcare organisations: Saudi healthcare as case study," *Comput Methods Programs Biomed*, vol. 212, 2021, doi: 10.1016/j.cmpb.2021.106466.

[9] F. Ayadi, "Critical factors affecting the decision to adopt cloud computing in Saudi health care organisations," *Electronic Journal of Information Systems in Developing Countries*, vol. 88, no. 6, 2022, doi: 10.1002/isd2.12231.

[10] F. Alharbi, A. Atkins, and C. Stanier, "Understanding the determinants of cloud computing adoption in Saudi healthcare organisations," *Complex & Intelligent Systems*, vol. 2, no. 3, pp. 155–171, 2016, doi: 10.1007/s40747-016-0021-9.

[11] L. G. Tornatzky, Mitchell. Fleischer, and A. K. Chakrabarti, *Processes of technological innovation*. Lexington Books, 1990.

[12] N. Zainuddin, N. Maarop, W. Z. Abidin, N. Firdaus Azmi, and G. N. Samy, "Cloud computing adoption conceptual model of Malaysian hospitals," *Open International Journal of Informatics (OIJI)*, vol. 3, no. 1, pp. 1–10, 2015.

[13] E. M. Rogers, *Diffusion of Innovations*, 4th ed. New York : Free Press, c1995., 1995.

[14] G. Agag and A. A. El-Masry, "Understanding consumer intention to participate in online travel community and effects on consumer intention to purchase travel online and WOM: An integration of innovation diffusion theory and TAM with trust," *Comput Human Behav*, vol. 60, pp. 97–111, 2016, doi: 10.1016/j.chb.2016.02.038.

[15] H. M. Sabi, F. M. E. Uzoka, and S. V. Mlay, "Staff perception towards cloud computing adoption at universities in a developing country," *Educ Inf Technol (Dordr)*, vol. 23, pp. 1825–1848, 2018, doi: 10.1007/s10639-018-9692-8.

[16] N. Alkhater, R. Walters, and G. Wills, "An empirical study of factors influencing cloud adoption among private sector organisations," *Telematics and Informatics*, vol. 35, no. 1, pp. 38–54, 2018, doi: 10.1016/j.tele.2017.09.017.

[17] A. N. Tashkandi and I. M. Al-Jabri, "Cloud computing adoption by higher education institutions in Saudi Arabia: An exploratory

study," *Cluster Comput*, vol. 18, no. 4, pp. 1527–1537, Dec. 2015, doi: 10.1007/s10586-015-0490-4.

[18] M. N. Birje, P. S. Challagidad, R. H. Goudar, and M. T. Tapale, "Cloud computing review: concepts, technology, challenges and security," *Int. J. Cloud Computing*, vol. 6, no. 1, pp. 32–57, 2017, doi: 10.1504/IJCC.2017.083905.

[19] A. Khayer, M. S. Talukder, Y. Bao, and M. N. Hossain, "Cloud computing adoption and its impact on SMEs' performance for cloud supported operations: A dual-stage analytical approach," *Technol Soc*, vol. 60, 2020, doi: 10.1016/j.techsoc.2019.101225.

[20] M. Mehrtak *et al.*, "Security challenges and solutions using healthcare cloud computing," *Journal of Medicine and Life*, vol. 14, no. 4. Carol Davila University Press, pp. 448–461, 2021. doi: 10.25122/jml-2021-0100.

[21] J. Opara-Martins, R. Sahandi, and F. Tian, "Critical analysis of vendor lock-in and its impact on cloud computing migration: a business perspective," *Journal of Cloud Computing*, vol. 5, no. 1, Dec. 2016, doi: 10.1186/s13677-016-0054-z.

[22] P. K. Yeng, S. D., and B. Yang, "Comparative Analysis of Threat Modeling Methods for Cloud Computing towards Healthcare Security Practice," *International Journal of Advanced Computer Science and Applications*, vol. 11, no. 11, pp. 772–784, 2020, doi: 10.14569/IJACSA.2020.0111194.

[23] D. Georgiou and C. Lambrinoudakis, "Compatibility of a security policy for a cloud-based healthcare system with the eu general data protection regulation (Gdpr)," *Information (Switzerland)*, vol. 11, no. 12, pp. 1–19, Dec. 2020, doi: 10.3390/info11120586.

[24] Y. Al-Issa, M. A. Ottom, and A. Tamrawi, "EHealth Cloud Security Challenges: A Survey," *Journal of Healthcare Engineering*, vol. 2019. Hindawi Limited, 2019. doi: 10.1155/2019/7516035.

[25] S. M. Gupta, "Cloud Security for Healthcare Services," *Journal of Management and Service Science (JMSS)*, vol. 3, no. 1, pp. 1–9, 2023, doi: 10.54060/jmss.v3i1.41.

[26] V. K. Prasad and M. D. Bhavsar, "Monitoring IaaS Cloud for Healthcare Systems: Healthcare Information Management and Cloud Resources Utilization," *International Journal of E-Health and Medical Communications*, vol. 11, no. 3, pp. 54–70, Jul. 2020, doi: 10.4018/IJEHMC.2020070104.

[27] M. Shahbaz and R. Zahid, "Probing the factors influencing cloud computing adoption in healthcare organisations: A three-way interaction model," *Technol Soc*, vol. 71, 2022, doi: 10.1016/j.techsoc.2022.102139.

[28] H. Gangwar, H. Date, and R. Ramaswamy, "Understanding determinants of cloud computing adoption using an integrated TAM-TOE model," *Journal of Enterprise Information Management*, vol. 28, no. 1, pp. 107–130, 2015, doi: 10.1108/JEIM-08-2013-0065.

[29] R. Martins, T. Oliveira, and M. A. Thomas, "An empirical analysis to assess the determinants of SaaS diffusion in firms," *Comput Human Behav*, vol. 62, pp. 19–33, 2016, doi: 10.1016/j.chb.2016.03.049.

[30] O. Ali, A. Shrestha, V. Osmanaj, and S. Muhammed, "Cloud computing technology adoption: an evaluation of key factors in local governments," *Information Technology and People*, vol. 34, no. 2, pp. 666–703, 2021, doi: 10.1108/ITP-03-2019-0119.

[31] C. Jianwen and K. Wakil, "A model for evaluating the vital factors affecting cloud computing adoption: Analysis of the services sector," *Kybernetes*, vol. 49, no. 10, pp. 2475–2492, 2020, doi: 10.1108/K-06-2019-0434.

[32] J. W. Lian, "Critical factors for cloud based e-invoice service adoption in Taiwan: An empirical study," *Int J Inf Manage*, vol. 35, no. 1, pp. 98–109, 2015, doi: 10.1016/j.ijinfomgt.2014.10.005.

[33] Z. Zafar, S. Islam, M. Shehzad, and M. Sohaib, "Cloud computing services for the healthcare industry," *International Journal of Multidisciplinary Sciences and Engineering (IJMSE)*, vol. 5, no. 7, pp. 25–29, 2014, [Online]. Available: www.ijmse.org

[34] O. Ali, A. Shrestha, J. Soar, and S. F. Wamba, "Cloud computing-enabled healthcare opportunities, issues, and applications: A systematic review," *Int J Inf Manage*, vol. 43, pp. 146–158, 2018, doi: 10.1016/j.ijinfomgt.2018.07.009.

[35] R. D. Raut, P. Priyadarshinee, B. B. Gardas, and M. K. Jha, "Analysing the factors influencing cloud computing adoption using three stage hybrid SEM-ANN-ISM (SEANIS) approach," *Technol Forecast Soc Change*, vol. 134, pp. 98–123, 2018, doi: 10.1016/j.techfore.2018.05.020.

[36] S. Midha *et al.*, "A Secure Multi-factor Authentication Protocol for Healthcare Services Using Cloud-based SDN," *Computers, Materials and Continua*, vol. 74, no. 2, pp. 3711–3726, 2023, doi: 10.32604/cmc.2023.027992.

[37] P. Khatiwada, H. Bhusal, A. Chatterjee, and M. W. Gerdes, "A Proposed Access Control-Based Privacy Preservation Model to Share Healthcare Data in Cloud," in *2020 16th International Conference on Wireless and Mobile Computing, Networking and Communications (WiMob)*, 2020, pp. 40–47. doi: 10.1109/WiMob50308.2020.9253414.

[38] M. Marwan, F. Sifou, F. AlShahwan, and A. A. Temghart, "An efficient privacy solution for electronic health records in cloud computing," *International Journal of High Performance Systems Architecture*, vol. 9, no. 4, pp. 201–214, 2020, doi: 10.1504/IJHPSA.2020.113681.

[39] N. Alkhater, G. Wills, and R. Walters, "Factors influencing an organisation's intention to adopt cloud computing in Saudi Arabia," in *2014 IEEE 6th International Conference on Cloud Computing Technology and Science*, IEEE Computer Society, 2014, pp. 1040–1044. doi: 10.1109/CloudCom.2014.95.

[40] E. AbuKhousa, N. Mohamed, and J. Al-Jaroodi, "E-Health cloud: Opportunities and challenges," *Future Internet*, vol. 4, no. 3, pp. 621–645, 2012, doi: 10.3390/fi4030621.

[41] A. Nirabi and S. A. Hameed, "Mobile cloud computing for emergency healthcare model: Framework," in *2018 7th International Conference on Computer and Communication Engineering (ICCCE)*, IEEE, 2018, pp. 375–379. doi: 10.1109/ICCCE.2018.8539310.

[42] M. Ijaz, G. Li, L. Lin, O. Cheikhrouhou, H. Hamam, and A. Noor, "Integration and applications of fog computing and cloud computing based on the internet of things for provision of healthcare services at home," *Electronics (Switzerland)*, vol. 10, no. 9, 2021, doi: 10.3390/electronics10091077.

[43] M. Mujinga, "Cloud computing inhibitors among small and medium enterprises," in *Proceeding of the 3rd International Conference on Intelligent Sustainable Systems, ICISS 2020*, IEEE, 2020, pp. 1385–1391. doi: 10.1109/ICISS49785.2020.9315905.

[44] M. Yuvaraj, "Perception of cloud computing in developing countries: A case study of Indian academic libraries," *Library Review*, vol. 65, no. 1–2, pp. 33–51, 2016, doi: 10.1108/LR-02-2015-0015.

[45] N. Masana and G. M. Muriithi, "Adoption of an integrated cloud-based electronic medical record system at public healthcare facilities in Free-State, South Africa," in *2019 Conference on Information Communications Technology and Society (ICTAS) 2019*, IEEE, 2019, pp. 1–6. doi: 10.1109/ICTAS.2019.8703606.

[46] S. A. Mokhtar, S. H. S. Ali, A. Al-Sharafi, and A. Aborujilah, "Organisational factors in the adoption of cloud computing in E-Learning," in *Proceedings - 3rd International Conference on Advanced Computer Science Applications and Technologies, ACSAT 2014*, IEEE, 2014, pp. 188–191. doi: 10.1109/ACSAT.2014.40.

[47] A. R. Stone and X. Zhang, "Understanding success factors for ERP implementation: An integration of literature and experience," *Issues In Information Systems*, vol. 22, no. 2, pp. 146–156, 2021, doi: 10.48009/2_iis_2021_150-161.

[48] S. Molinillo and A. Japutra, "Organisational adoption of digital information and technology: A theoretical review," *Bottom Line*, vol. 30, no. 1, pp. 33–46, 2017, doi: 10.1108/BL-01-2017-0002.

[49] S. Chaudhry, "Managing employee attitude for a successful information system implementation: A change management perspective," *Journal of International Technology and Information Management*, vol. 27, no. 1, 2018, doi: 10.58729/1941-6679.1364.

[50] E. Toufaily, T. Zalan, and S. Ben Dhaou, "A framework of blockchain technology adoption: An investigation of challenges and expected value," *Information and Management*, vol. 58, no. 3, 2021, doi: 10.1016/j.im.2021.103444.

[51] X. Wang, L. Liu, J. Liu, and X. Huang, "Understanding the determinants of blockchain technology adoption in the

construction industry," *Buildings*, vol. 12, no. 10, 2022, doi: 10.3390/buildings12101709.

[52] C. Changchit and C. Chuchuen, "Cloud computing: An examination of factors impacting users' adoption," *Journal of Computer Information Systems*, vol. 58, no. 1, pp. 1–9, 2018, doi: 10.1080/08874417.2016.1180651.

[53] R. Vidhyalakshmi and V. Kumar, "Determinants of cloud computing adoption by SMEs," *Int J Bus Inf Syst*, vol. 22, no. 3, pp. 375–395, 2016, doi: 10.1504/IJBIS.2016.076878.

[54] V. Rai *et al.*, "Cloud Computing in Healthcare Industries: Opportunities and Challenges," in *Innovations in Computing*, Singapore: Springer, Apr. 2022, pp. 695–707. doi: 10.1007/978-981-16-8892-8_53.

[55] S. Asif, M. Ambreen, Z. Muhammad, H. ur Rahman, and S. Iqbal, "Cloud Computing in Healthcare - Investigation of Threats, Vulnerabilities, Future Challenges and Counter Measure," *LC International Journal of STEM*, vol. 3, no. 1, pp. 63–74, 2022, doi: 10.5281/zenodo.6547289.

[56] Y. A. M. Qasem *et al.*, "A multi-analytical approach to predict the determinants of cloud computing adoption in higher education institutions," *Applied Sciences (Switzerland)*, vol. 10, no. 14, 2020, doi: 10.3390/app10144905.

[57] I. Ahmed, "Technology organisation environment framework in cloud computing," *Telkomnika (Telecommunication Computing Electronics and Control)*, vol. 18, no. 2, pp. 716–725, 2020, doi: 10.12928/TELKOMNIKA.v18i2.13871.

[58] K. K. Hiran and A. Henten, "An integrated TOE–DoI framework for cloud computing adoption in the higher education sector: case study of Sub-Saharan Africa, Ethiopia," *International Journal of System Assurance Engineering and Management*, vol. 11, no. 2, pp. 441–449, 2020, doi: 10.1007/s13198-019-00872-z.

[59] A. Khoirunnida, A. N. Hidayanto, B. Purwandari, D. Kartika, and M. Kosandi, "Factors influencing citizen's intention to participate electronically: The perspectives of social cognitive theory and e-government service quality," in *2017 International Conference on Advanced Computer Science and Information Systems (ICACSIS)*, 2017, pp. 166–171. doi: 10.1109/ICACSIS.2017.8355028.

[60] A. Gutierrez, E. Boukrami, and R. Lumsden, "Technological, organisational and environmental factors influencing managers' decision to adopt cloud computing in the UK," *Journal of Enterprise Information Management*, vol. 28, no. 6, pp. 788–807, 2015, doi: 10.1108/JEIM-01-2015-0001.

[61] J. Sun, "Tool choice in innovation diffusion: A human activity readiness theory," *Comput Human Behav*, vol. 59, pp. 283–294, 2016, doi: 10.1016/j.chb.2016.02.014.

[62] M. Zhou, R. Zhang, W. Xie, W. Qian, and A. Zhou, "Security and privacy in cloud computing: A survey," in *Proceedings - 6th International Conference on Semantics, Knowledge and Grid, SKG 2010*, 2010, pp. 105–112. doi: 10.1109/SKG.2010.19.

[63] H. Gupta and D. Kumar, "Security threats in cloud computing," in *Proceedings of the International Conference on Intelligent Computing and Control Systems (ICICCS 2019)*, IEEE, 2019, pp. 1158–1162. doi: 10.1109/ICCS45141.2019.9065542.

[64] Z. Mahmood, "Data location and security issues in cloud computing," in *Proceedings - 2011 International Conference on Emerging Intelligent Data and Web Technologies, EIDWT 2011*, 2011, pp. 49–54. doi: 10.1109/EIDWT.2011.16.

[65] A. Rashidi and N. Movahhedinia, "A model for user trust in cloud computing," *International Journal on Cloud Computing: Services and Architecture*, vol. 2, no. 2, pp. 1–8, 2012, doi: 10.5121/ijccsa.2012.2201.

[66] N. M. Sultana and K. Srinivas, "Survey on centric data protection method for cloud storage application," in *2021 International Conference on Computational Intelligence and Computing Applications, ICCICA 2021*, IEEE, 2021, pp. 1–8. doi: 10.1109/ICCICA52458.2021.9697235.

[67] S. Sengupta, V. Kaulgud, and V. S. Sharma, "Cloud computing security--trends and research directions," in *2011 IEEE World Congress on Services*, IEEE, 2011, pp. 524–531. doi: 10.1109/services.2011.20.

[68] M. A. Alanezi, "Factors influencing cloud computing adoption in Saudi Arabia's private and public organisations: A qualitative evaluation," *International Journal of Advanced Computer Science and Applications (IJACSA)* , vol. 9, no. 4, pp. 121–129, 2018, [Online]. Available: www.ijacsa.thesai.org

[69] M. O. Alassafi, R. Alghamdi, A. Alshdadi, A. al Abdulwahid, and S. T. Bakhsh, "Determining factors pertaining to cloud security adoption framework in government organisations: An exploratory study," *IEEE Access*, vol. 7, pp. 136822–136835, 2019, doi: 10.1109/ACCESS.2019.2942424.

[70] M. Masrom and A. Rahimli, "Cloud computing adoption in the healthcare sector: A SWOT analysis," *Asian Soc Sci*, vol. 11, no. 10, pp. 12–18, 2015, doi: 10.5539/ass.v11n10p12.

[71] F. Mohammed, O. Ibrahim, and N. Ithnin, "Factors influencing cloud computing adoption for e-government implementation in developing countries: Instrument development," *Journal of Systems and Information Technology*, vol. 18, no. 3, pp. 297–327, 2016, doi: 10.1108/JSIT-01-2016-0001.

[72] A. D. Abubakar, J. M. Bass, and I. Allison, "Cloud computing: Adoption issues for sub-saharan African SMEs," *Electronic Journal of Information Systems in Developing Countries*, vol. 62, no. 1, pp. 1–17, 2014, doi: 10.1002/j.1681-4835.2014.tb00439.x.

[73] M. Odeh, A. Garcia-Perez, and K. Warwick, "Cloud computing adoption at higher education institutions in developing countries: A qualitative investigation of main enablers and barriers," *International Journal of Information and Education Technology*, vol. 7, no. 12, pp. 921–927, 2017, doi: 10.18178/ijiet.2017.7.12.996.

[74] F. Alharbi, A. Atkins, and C. Stanier, "Decision makers views of factors affecting cloud computing adoption in saudi healthcare organisations," in *2017 International Conference on Informatics, Health and Technology, ICIHT 2017*, IEEE, 2017, pp. 1–8. doi: 10.1109/ICIHT.2017.7899001.

[75] M. M. Lawan, C. F. Oduoza, and K. Buckley, "Proposing a conceptual model for cloud computing adoption in upstream oil & gas sector," *Procedia Manuf*, vol. 51, pp. 953–959, 2020, doi: 10.1016/j.promfg.2020.10.134.

[76] M. Al-Ruithe, E. Benkhelifa, and K. Hameed, "Current state of cloud computing adoption - An empirical study in major public sector organisations of Saudi Arabia (KSA)," *Procedia Comput Sci*, vol. 110, pp. 378–385, 2017, doi: 10.1016/j.procs.2017.06.080.

[77] MCIT Reports, "KSA Cloud First Policy," 2020. [Online]. Available: https://www.mcit.gov.sa/sites/default/files/cloud_policy_en.pdf

[78] V. Venkatesh, S. A. Brown, and H. Bala, "Bridging the qualitative-quantitative divide: Guidelines for conducting mixed methods research in information systems," *MIS Quarterly*, vol. 37, no. 1, pp. 21–54, 2013, [Online]. Available: http://www.jstor.org/stable/43825936

Position Papers of the 18th Conference on Computer Science and Intelligence Systems pp. 27–34
DOI: 10.15439/2023F3591
ISSN 2300-5963 ACSIS, Vol. 36

An Outlook on Natural Language Generation

Anabela Barreiro
0000-0001-9521-3006
INESC-ID, Lisboa
Rua Alves Redol, 9
1000-029 Lisboa, Portugal
anabela.barreiro@inesc-id.pt

Elena Lloret
0000-0002-2926-294X
Universidad de Alicante
ctra. San Vicente s/n
03690 San Vicente, Alicante, Spain
elloret@dlsi.ua.es

Oleksii Turuta
0000-0002-0970-8617
Kharkiv N. Univ. of Radio Electrics
Nauky Ave. 14,
61165 Kharkiv, Ukraine
oleksii.turuta@nure.ua

***Abstract*—This article presents an outlook on the present state of Natural Language Generation, discusses its impact and research challenges in light of recent progress in the development of Large Language Models. We foresee adverse results and repercussions arising from the use of models like ChatGPT if they are applied inappropriately in the absence of adequate legal guidelines to regulate their usage. Our aim is to highlight the importance of safeguarding human rights and provide recommendations for addressing the most pressing issues to ensure the long-term viability and security of this technology. In the context of the Multi3Generation COST Action (CA18231), we suggest strategies to address intricate challenges and showcase ongoing projects in strategically important European regions: (1) the long-established Iberian Peninsula, home to two of the world's most widely spoken languages, along with several minority languages, and (2) Ukraine, which is actively pursuing its right to self-determination, autonomy, and the preservation of its cultural and linguistic identity. We aim at developing and strengthening a common strategy for newer models that can work alongside or in conjunction with existing ones. The main difference we wish to make in research is to focus on the science of language, further exploring linguistic resources and using their fine-grain quality to create new systems and enhance existing ones in a trusting atmosphere between developers and users, and a safe and innovation-friendly environment for society at large.**

I. Introduction

AS WORDS have become almost meaningless in mass and social media, texts produced massively have become increasingly trivial. Parallel to this, the quality of data and the fidelity, credibility, truthfulness, and trustworthiness of the information conveyed to the masses, sometimes *ad nausea*, is incrementally questionable in terms of false assertions and veracity of information with relation to facts. This appears to be a critical time for reflection on the future of natural language generation (NLG) technology, its methods, models, and applications. Our debate comes at a time of multiple discussions around the widespread disruption brought by large language models (LLMs), particularly followed by the most recent releases of ChatGPT by OpenAI and models alike. This progress may bring the idea that recent models represent the industry's endgame, which causes fear in some, and awe in others, especially due to their enormous power to be used for good and for evil in many diverse and varied forms. Therefore, the good results of the new models should be viewed with some reservation and a critical eye over its developments. On the one hand, the exponential capabilities of LLMs and their capacity to learn from vast datasets, primarily comprised of collective human knowledge present a tremendous technological opportunity, as they are evolving into a multimodal direction, not limited to processing text alone but also incorporating elements like audio-visual data (including colors, images, videos, speech, etc.), addressing complex and high-dimensional phenomena, and seamless managing multi-tasking demands. On the other hand, the increasing reliance on powerful machine learning (ML) algorithms in AI has the potential to push human control to its limits, particularly in domains like education, medical science, business operations, and most critically, space and military applications.

Multi3Generation COST Action (CA18231), henceforth simply Multi3Generation[1] [**?**], is a European Excellence Network that focuses on NLG research field from a very broad perspective, ranging from machine learning methods, including LLMs to more linguistic methodologies. Through the network exchange facilitated by Multi3Generation, involving researchers from 33 countries, we seize the opportunity to propose alternative or complementary methods for progressing in the field of NLG by promoting the development of (semi- or partially-)transparent glass box systems, instead of blurry or black box systems, i.e., systems built where the human in the loop is essential in the process of language generation, and most importantly, systems that move towards language understanding because they contain knowledge embedded in them, not systems which use questionable data, from questionable sources and questionable methods of obtaining, processing and disseminating those data. Beyond these considerations, we wish to emphasize that we acknowledge the significance of advancements in models that simulate/mimic language, which are already recognized as a noteworthy achievement with substantial societal impact. However, we mostly want to reinforce the opportunity for humans to unravel and grow their consciousness of what they can do better than machines. We also highlight the need to explore innovative approaches and flexible architectures, as well as the adoption of hybrid strategies, which use parts of different approaches and methodologies to the design of 'safe' systems, especially those that make progress in the understanding of language as a science, an achievement that is far from being accomplished up to now.

[1] https://multi3generation.eu/

©2023, PTI

Thematic track: Challenges for Natural Language Processing

The remainder of the article is organized as follows. Section II presents state-of-the-art NLG, focusing on current projects being developed in the Iberian Peninsula in line with the Portuguese and Spanish strategies for AI, and also emphasize the efforts undergoing in Ukraine to combat misinformation and Russian propaganda in the media. Section III discusses some AI ethical issues that, in our view, require discussion among researchers before they get out of control in the real world. Section IV focuses on the initiatives taken within Multi3Generation to promote innovation, foster interdisciplinary understanding, and encourage the sustainable development of NLG and responsible AI. Finally, Section V presents the conclusions and future work.

II. State of the Art

Natural Language Generation (NLG) has a well-established research history with a track record of success dating back to the early days of natural language processing (NLP) systems, particularly machine translation [**?**], [**?**], [**?**], [**?**]. Some projects have been developed, drawing inspiration from the Logos Model's language generation capabilities. These projects have explored linguistics and applied them to generate paraphrases[2] [**?**], [**?**], conducting experiments in various areas, including translation [**?**], language varieties [**?**], [**?**], [**?**], stylistics [**?**], emotions [**?**], question-answering, and summarization tasks [**?**]. These efforts have led to the identification of a new linguistic corollary concept known as a "paraphrasary" [**?**], [**?**], result of the experiments conducted alongside the primary paraphrasing research. The research has been extensively documented and is fully traceable.

Due to a scarcity of comprehensive documentation enabling efficient traceability, the sudden rise of generative LLMs took many individuals by surprise. These models began exhibiting generative capabilities that could be both overwhelming and/or controversial for researchers, developers, and users alike. Humans have long been aware that a small gadget has the information and the ability to grow the amount of information of an encyclopedia, an entire library, an entire university, or the world of universities altogether. Now, humans are becoming aware that machines may produce texts of a better quality than them. More and more, digital devices are indispensable to help us in many tasks, and there is a strong human-machine interaction in our daily lives. The technology is here to stay and help. Many ongoing research projects are using LLMs, either exploring their potential, as well as addressing some of their limitations. Here, we focus on projects under development in the Iberian Peninsula and efforts carried out in Ukraine, which we find relevant for further development of AI with regards to NLG. In these projects, there is an implicit intention to distinguish and complement themselves from already existing surveys of Anglo-Saxon NLG solutions [**?**] [**?**], and research and commercial state-of-the-art articles [**?**] [**?**]. We believe that, if responsibly implemented, new hybrid models with a social and linguistic motivation behind them can be more trustful, controllable, and efficient in the long run. The projects described here have in common a sense of social responsibility, inclusiveness, ethical standards and responsible care in their design, even when LLMs play a role in them. Members of the Multi3Generation network participate and/or have been partially funded by Multi3Generation.

In Portugal, INESC-ID[3] is a center of excellence for AI and is currently involved in cutting-edge research in AI bridging the gap between academia and research. The interdisciplinarity exchange among researchers working in AI for people and society, information and decision support systems, and human language technologies have been surrounding R&D fields that converge in the area of NLG. We can point out 3 ongoing NLG-related projects: (i) The Center for Responsible AI (CRAI), (ii) Accelerating digital transformation in Portugal (Accelerat.ai), and (iii) The Multimodal Approach for Identifying Conspiracy Theories in Social Media (MAICT). In the next paragraphs, we will describe in detail each one of them.

Financed by the European Union, the "Center for Responsible AI" (CRAI)[4] operates as a component of the Recovery and Resilience Plan (PRR). Headed by the Portuguese startup Unbabel, and other 9 startups (including two highly valued unicorns), in coordination with INESC-ID and 7 other research institutes, the CRAI Center stands out as one of the most extensive initiatives with a focus on ethics and responsibility in AI. The consortium includes a legal firm. Through collaborative efforts among these partners, the aim is to create 21 groundbreaking AI solutions/products that integrate responsible/ethical AI principles such as equity, explainability, and sustainability. This will not only position Portugal and Europe as global front-runners in these technologies but will also help establish guiding principles and regulations within the realm of AI. The joint collaboration in CRAI will play a pivotal role in shaping European Union legislation concerning AI and in attracting international top-tier AI talent.

INESC-ID is also a partner in Accelerat.ai[5], a new large consortium with the objective of expediting the digital evolution of both public and private sectors in Portugal. This project is also backed by the Portuguese PRR, led by the startup Defined.ai and partnered with several corporations and research institutes. Accelerat.ai's main aim is to enhance and optimize customer support services in the Portuguese and European markets, based on the development of a unique set of technological solutions that combine virtual assistants with European Portuguese contact centers, and introduce Portugal to the first fully native-language virtual assistant. INESC-ID's participation in Accelerat.ai, focuses primarily on investigating and exploring the capability of mutual conversion between speech and text, commonly referred to as Automatic Speech Recognition and Speech Synthesis, essential components of Conversational AI. INESC-ID and IST are taking charge of de-

[2]https://www.inesc-id.pt/projects/IP02043/

[3]https://www.inesc-id.pt/

[4]https://www.inesc-id.pt/inesc-id-takes-part-in-the-worlds-largest-consortium-on-responsible-ai/

[5]https://www.inesc-id.pt/accelerate-ai-a-new-consortium-to-accelerate-digital-transformation-in-portugal/

veloping and researching modules within the speech-to-speech conversational framework. This emphasis includes areas like automatic speech recognition and text-to-speech synthesis for European languages, especially European Portuguese. The expected impact of this research lies in facilitating the fast deployment of new languages, creating more engaging and successful agents, and enhancing acceptance among specific demographic groups.

The exploratory project "Multimodal Approach for Identifying Conspiracy Theories in Social Media" (MAICT)[6], led by INESC-ID, and funded by the Portuguese government (FCT project EXPL/LLT-LIN/1104/2021) aims at developing an innovative multimodal (text-image) conspiracy grammar for Portuguese, based on a multidisciplinary approach. The project contributes to advancing the state-of-the-art in multimodal misinformation studies and addresses the following research questions: 1) Which are the most predominant morphosyntactic, lexico-syntactic, semantic, and discursive features in conspiracy theory narratives that abound in social media, and how do they relate to each other? (2) Which meanings can be inferred from images that include conspiracy narratives, from a social semiotics point of view? 3) How do text and image articulate in multimodal conspiracy narratives to create either a unified or a dissociated meaning? 4) Which are the most suitable approaches to describe and formalize the multimodal properties from text and image, and respective interaction, in view of their automatic processing? 5) How can Portuguese conspiracy sociolects be characterized from a multimodal point of view? Some of the outcomes of this project can certainly enrich emerging transparent models envisaged in Multi3Generation.

The Spanish government is committed to supporting strategic projects for responsible AI, focusing on thematic areas such as employment of commonsense reasoning, inclusion via the use of accessible and intelligible content, improvement of fake news detection, and empowerment of languages with scarce resources. We mention here 4 projects in which the University of Alicante is enrolled: (i) the NLG-related project CORTEX that aims to enhance the commonsense reasoning competence of NLG systems; (ii) the inclusive CLEAR.TEXT project that explores resilient technologies to assist in creating accessible content, (iii) the NL4DISMIS project that aims to identify false and incorrect information in texts and use NLG and/or automated methods to either counteract or supplement the misleading information; and (iv) the NEL-VIVES collaborative project among various Spanish institutions hailing from distinct regions focusing particularly on analyzing and advancing LLMs for Spanish official languages with limited resources. Next, we will describe in further detail each one of these projects.

The "Conscious Text Generation" (CORTEX)[7] is an R&D project funded by the Spanish government, and by "ERDF A way of making Europe", that deals with the integration of world and external knowledge in NLG architectures in order to improve the commonsense reasoning capabilities of NLG systems. The project considers that enhancing the commonsense reasoning capabilities of NLG systems is needed to automatically produce accurate, correct, and reliable texts that will be in line with real facts. The main research questions proposed in this project are: 1) How to address/mitigate the problem of hallucination? 2) How to ensure the provided information is reliable?

The project "Enhancing the modernization public sector organizations by deploying NLP to make their digital content CLEARER to those with cognitive disabilities" (CLEAR.TEXT)[8] is funded by the Spanish government and the European Union and its objective is to research, implement, deploy, evaluate, and ultimately provide robust technologies for NLP to support the authoring of accessible Spanish content for public sector organizations (at the local, regional and national level) that is intelligible to people with a cognitive disability, thereby widening their inclusion and empowerment in Europe. The research question that is being investigated is: 1) How to make the information clearer depending on the user's needs? For this, the project analyses the capabilities of LLMs, and derived tools, such as ChatGPT to generate automatic simplified summaries.

The project "Natural Language Technologies for dealing with dis- and misinformation" (NL4DISMIS)[9] is funded by the Generalitat Valenciana. The main hypothesis of this project is based on the existence of a direct relation between the use of human language (i.e., language models) and the user's behavior in digital media. Therefore, by modeling the language used within a contextualization at different linguistic levels, we can establish the relation between different entities, as well as the evolution of these entities and their relations over time. Simultaneously, it may be possible to infer new relations and predict future states or behaviors. The evolution of entities over time requires research on entity language models as well as knowledge representation based on digital entities. The main objective of the project is to detect dis- and misinformation in texts and automatically debunk misleading information through NLG and/or complement the information automatically. The research question to address is: 1) How to ensure the provided information is reliable?

The project "Language Technologies Plan for Valencian language" (NEL-VIVES)[10] is funded by PERTE Nueva Economía de la Lengua from the Spanish Government. This is a coordinated project between several Spanish intuitions from different regions (Catalonia, Basque Country, Galicia, and Valencian Region) with emphasis on the analysis and development of LLMs for low-resource Spanish official languages. The research question under this project is: 1) How can we preserve and promote languages in danger of digital extinction? In

[6]https://maict.inesc-id.pt/

[7]MCIN/AEI/10.13039/501100011033/ – PID2021-123956OB-I00 https://cortex.gplsi.es/en/home/

[8]MCIN/AEI/10.13039/501100011033 NextGenerationEU/PRTR – TED2021-130707B-I00 https://cleartext.gplsi.es/en/home/

[9]CIPROM/2021/21 – https://nl4dismis.gplsi.es/

[10]https://vives.gplsi.es/

this sense, the subproject NEL-VIVES focuses on the Valencian language, in particular, the goal is to create corpora (text+voice) and develop language models for the different varieties of Valencian language.

In Ukraine, the coordination of AI development is entrusted to the Ministry of Digital Transformation. An advisory body, the AI Committee, has been established, bringing together researchers, businessmen, and policy makers. Ukrainian universities such as Kharkiv National University of Radio Electronics (NURE) and Ukrainian Catholic University (UCU) are independently advancing NLP technologies. To support startups, the Ukrainian Startup Fund[11], a state-funded initiative, has been established in 2020. Within the WG1 of Multi3Generation, a multimodal dataset, Multi30K-UK[12], has been created, containing images with descriptions in Ukrainian. In the conditions of Russia's military aggression against Ukraine, startups focused on detecting misinformation are emerging. For example, the AI startup Osavul has developed technology to combat propaganda and Russian disinformation, assisting the National Security and Defense Council (RNBO) and the Ministry of Defense in detecting information warfare activities on platforms like Telegram and Facebook.[13]

A common aspect of all the projects outlined here (the list of ongoing R&D projects in the Iberian Peninsula and those under development in Ukraine) is that they have the common ground of being inclusive, and interdisciplinary, crossing knowledge and methods from AI such as human-centered AI, computational models of narrative and discourse, neurosymbolism, cognitive vision, space, speech and language technologies and NLP areas such as cognitive linguistics, corpus linguistics, forensic linguistics, social semiotics, among others. But, mostly, they have the main interest in creating or generating responsible/ethical LLMs or other language models. In Section III, we lay out some particular concerns related to LLMs, namely ChatGPT.

III. Ethics in Natural Language Generation

GPT stands for "Generative Pre-trained Transformer", and it follows older models that learn statistical regularities in language to a greater or lesser extent.[14] They synthesize existing content to make it appear to be new content by the power that the system has of recognizing and learning statistical regularities in language, inducting from information already stored in the system or information provided by new interactions with the user who provides the system with that information, and lately, predicting language patterns and escalating the use of its language resources in order to improve its performance. In this sense, the so-called generative language does not involve creativity, but rather different degrees of 'concealed regurgitation' facilitated by smart algorithms that allow the recognition of patterns but are incapable of understanding natural human language in the way humans do. In addition, LLMs are trained with zillions and zillions of data that are only accessible to very few who can afford to use the data created by everyone. As far as the use of AI for creative purposes, initiatives such as "AI, Generation and Creativity" (AIGC) give rise to the AIGC models used and trained by informed users to create their own unique content generation models [**?**]. The technology creates content based on algorithms, models, and rules, but does not dispense the user as the provider of data, as its content is generated based on user-inputted keywords or requirements via crowdsourcing techniques or others. User-generated content is increasingly influencing editorial choices of content. AI tools can be a supplement rather than a replacement for human creativity and it is up to humans to find the balance between efficiency and creativity.

We wish to emphasize that we are not against LLMs and acknowledge their potential and usefulness to the human being. However, we ought to stress that this needs to be done responsibly and respect the Code of Ethical Conduct. No nation or corporation should have the power and control over the most fundamental inalienable rights of individuals, especially when it concerns their unique linguistic, and cultural identity, as well as the core values. Language is an integral part of our human identity and the most direct expression of our culture, which are sovereign assets of their native speakers. We believe that it is essential to debate on international laws to protect the collective identity of citizens. By our promotion of a deep study of languages, we imply that linguistic values should be safeguarded. Furthermore, we hold the belief that while Large Language Models (LLMs) can be beneficial, they may not necessarily represent the ideal solution for numerous tasks or objectives.

There are reasonable concerns over misuse, leaks in data protection, missing regulations related to confidentiality, and lack of anonymization of individual records, among others, as admitted by the men at the cutting edge of the technology. Faced with the dangers brought by the power of the technology that can be used in the wrong direction, it urges the adoption of legislation foreseeing authenticity, security, reliability, and integrity of (proprietary) information. To put it bluntly, without creating the mechanisms necessary to minimize dangerous risks and their societal impact, a 'blind' adoption of LLMs in NLP tasks including fully automatic high-quality machine translation (FAHQMT), summarisation, and text simplification should be still considered an illusion comparable to the hallucinations that appear here and there in the outputs of LLM-based systems. As a matter of fact, society needs to know if/when it can rely or not rely on the current technology, and be aware of the implications in case of misuse or overlooking of the technology's flaws. Even if the outputs of the new models appear often better than those produced by humans, no one should be 'fooled' by a machine that does not have a conscience of what is generated when the content

[11] https://usf.com.ua/

[12] https://aclanthology.org/2023.unlp-1.7

[13] www.osavul.cloud/ai-against-russian-ipso-ukrainian-startup-osavul-taught-neural-networks-to-fight-propaganda-how-to-sell-such-technology

[14] They also use neural networks on data, but it is not relevant for this point of discussion.

produced is not totally reliable. In our view, effective linguistic quality control and fact-checking (among others) should be collectively addressed by professional experts. Currently, with a few exceptions [**?**], there is a lack of scientific evaluation for the new NLG technologies.

Section IV presents a series of initiatives within Multi3Generation that aim to establish a network of researchers who could have an interest in participating in the development and evaluation of resources of the type just illustrated or in combining these resources in hybrid systems or developing improved and controlled NLG models.

IV. Multi3Generation Initiatives

Multi3Generation gathers a network of researchers who work on the progress of multilingual, multimodal, and multi-task NLG. Within the scope of the Action, several initiatives can be emphasized: Short Term Scientific Missions (STSMs), Training Schools (TS); and specific workshops. In IV-A, IV-B, and IV-C, we present the outcomes of the activities accomplished up to now.

A. Short Term Scientific Missions

Short Term Scientific Missions allow researchers to visit groups and institutions located in countries participating in Multi3Generation to create synergies as well as to improve their research abilities. It is an efficient mechanism to promote joint collaborations as well as to disseminate research in NLG among different countries. From the beginning of the Action up to now, 20 STSMs have been completed[15], all of them focusing on the NLG topics. It is worth mentioning them here because they focused on exploring NLG and its applications to sectors, such as education as well as those addressing the improvement of NLG by integrating knowledge, multimodal information, or analyzing more efficient methods. The topics were: (1) multi-task sequence learning for syntax; (2) analysis and introspection of multilingual representations; (3) image captioning using relational context from generated scene graphs; (4) fusion mechanisms in claim verification models; (5) natural language grounding; (6) generating fact-checking explanations in low-resource settings; (7) morphological typology awareness in multilingual NLP evaluation; (8) enhancing NLG with knowledge acquisition/integration; (9) exploring the interplay between grammatical and cultural gender for debiased NLG; (10) text summarization as a digital tool to be applied in writing for academic purposes; (11) investigating the discrepancy in probing techniques for verb understanding in image-language transformers; (12) challenges and obstacles zero-shot multimodal reasoning with language; (13) Graph2Seq models for NLG tasks; (14) NLG and text summarization: exploring use cases in education; (15) using knowledge graphs to improve NLG tasks; (16) multimodal interactions in the collaborative industrial environment – empirical and analytical methods; (17) generating code from multilingual prompts; (18) fusion of multimedia information in deep learning models; (19) digitalization of humanities; (20) counting repetitive actions using pretrained video-and-language models. Some of these STSMs can result into research articles or transnational collaborative projects.

B. Training Schools

Multi3Generation organized two training schools and is preparing a third one. These training schools were designed to attract young, early career academics in specific topics of the COST Action. The training schools address NLG challenges in the digital realm, specifically in human-machine interaction in selected application areas of emerging societal significance, such as language technologies, which affect EU citizens in an accessible, multilingual Europe. They also target professional and recreational needs that will have strong economic and societal impacts.

1) Creative Natural Language Generation: The training school on Creative Natural Language Generation[16] brought together experts on computationally-oriented methodologies with experts on theories and insights from the humanities (computational linguistics, psychology, media studies, philological and literary studies, etc.). From this interdisciplinarity and clash of ideas, students could (i) learn different theories to construct metaphorical expressions using affection, persuasive and even humorous language, visualize some pilot computational experiments on typical creative genres, and judge the level/type of creativity of an AI metaphor generator; (ii) learn what rule-based automatic text generation (ATG) task is, and explore how to generate text capturing the expressivity of natural languages in machine-representation systems, such as knowledge bases, taxonomies, and ontologies, namely grasp the difficult task of writing automatically the basic plot of a novel; (iii) learn to develop linguistic resources for NLG using NooJ; (iv) get acquainted with techniques to annotate sentiment and emotions in literary texts, and learn how to create and use domain-specific languages (DSLs) to annotate literary texts, with benefits such as compactness, familiarity, and completeness, among others. Students also were offered a step-by-step tutorial on context-free grammar and annotated empathic expressions through a domain-specific language called EmpathyDSL, created for the specific task; (v) have an overview of multilingual language resources and NLP tools and services available in CLARIN, discover new resources, deposit and preserve newly created ones, find tools that can process and annotate them, and test important CLARIN services, such as the Virtual Language Observatory and Switchboard; (vi) reflect on the many facets of affective expressions in multilingual text; (vii) generate comics with meaning, intent, and humor, which integrate words and images to support more possibilities than either a text or an image alone can offer and imbue events with strong emotions. This rich combination of research topics will certainly turn into interesting, avant-garde projects.

[15]https://multi3generation.eu/funding-opportunities/short-term-scientific-missions/

[16]https://multi3generation.eu/2022/06/24/m3g-cost-action-training-school-on-creative-natural-language-generation/

2) Representation Mediated Multimodality: Undoubtedly, multimodality is a key issue within NLG, as more and more inputs go beyond texts, also including video, images, audio, etc. The training school on Representation Mediated Multimodality[17] aimed to provide a joint perspective on the theoretical, methodological and applied understanding of representation-mediated multimodal sense-making at the interface of language, knowledge representation and reasoning, and visuo-auditory computing. A primary topic developed in this TS was grounded meaning-making. Grounding is a challenge for NLG since it has to do with the semiotic construction of language. Linguistic expressions and/or relational categorizations are called grounded when they are linked to non-linguistic, especially quantitative perceptual data, such as information coming from modalities such as vision and audition in space-time. Such perceptual data could pertain to, for instance, dynamic spatio-temporal phenomena both in an embodied as well as disembodied interaction context. Grounding is, in essence, a key aspect of semiotic construction, e.g., enabling high-level meaning acquisition, and analogy, and has been a long-standing challenge in AI and related disciplines. Within the central topic of grounding, other topics covered in this TS were: explainable multimodal commonsense understanding, multimodal generation/synthesis for communication, multimodal summarization, multimodal interpretation-guided decision-support, adaptation & autonomy, and analytical visualization. The topics covered benefited not only researchers and experts in the NLG field but anyone interested in human-machine interaction and NLP, as the contents of the TS have a direct application in both public and private human-centered technological services.

C. Workshops and other events

Multi3Generation was represented in workshops or conferences (invited participant and/or organizer) listed here: (1) the 6th conference of the European Language Resource Coordination.[18] The talk focused on "Multi3Generation – Multimodal Data for Natural Language Generation: Current Contributions and Future Perspectives" of Multi3Generation in the topic Think BIG. For Europe's Multilingual FUTURE; (2) the 13th International Conference on Natural Language Generation, Dublin, Ireland[19]; (3) the round table "The European panorama of linguistic technologies"[20] of the summer course at the Escola Técnica Superior de Enxeñaría @ Campus Vida of the University of Santiago de Compostela; (4) the Spanish online round table "NLP in International Projects" for the DiverTLes community, Meeting of Natural Language Processing Companies, a forum organized by PERTE Nueva economía de la lengua[21]; (5) the online colloquium "In Translation"[22] organized by the Institute of Humanities, Faculty of Human and Social Studies, Mykolas Romeris University, Vilnius, Lithuania. The talk's topic was "Linguistic resources for the translation of creative language"; (6) the 23rd Annual Conference of the European Association for Machine Translation with a meta-paper on the COST Action; (7) the organization of the workshop: COST Action Multilingual, Multimodal and Multitask Language Generation, co-located with the 24th Annual Conference of The European Association for Machine Translation, which took place in Tampere, Finland.[23] Some of the participation in these events resulted in publications.[24]

Multi3Generation continues to provide the opportunity for professional training of young people through the reinforcement of digital skills at all levels of qualification and teaching and training modalities. In addition, Multi3Generation promotes the improvement of NLG systems using knowledge-enhanced approaches. It notes that LLMs based on Transformers can be adjusted and fine-tuned to enhance accuracy and control in text generation. These adjustments would result in more natural, varied text, while also addressing issues like biases and misleading content. This level of control is essential for effectively applying NLG systems in real-world contexts within industry and society.

V. Conclusions and Future Work

This paper presented an outlook on current NLG, highlighted some concerns with regard to the use of LLMs and outlined the need to complementary approaches. After the announcement of the great leap forward of AI, it is time for a deep thorough reflection of the implications of AI in the life of humans in general. In the context of Multi3Generation, we put forth a range of initiatives aimed at addressing intricate challenges in the NLG field. Additionally, we present ongoing projects located in two distinct European regions: the western corner, represented by the Iberian Peninsula (Portugal and Sapin), and the Eastern European region of Ukraine. The primary objective is to develop and improve a collaborative strategy for emerging models. Many projects will result from the release of OpenAI in multidisciplinary areas and topics. Our main concern is: will and should AI be possible without including the science of language in all its sub-fields (psycholinguistics, social linguistics, among others)? The main reason for this question is related to the fact that linguistic-based methods have been left behind in the field/history and discussions of AI as if the processing of language is a task that can be done without linguists. It appears to us that, as a science, linguistics has been 'cantoned' from language models for decades. It has not been given the importance that it has or the role it must play. There is also a great discrepancy with regard to the attention that some languages have been given in comparison to others. While English is the most processed language of all for the reason that it works as "*lingua franca*", some languages still lack the amount and quality of

[17] https://multi3generation.eu/2022/05/05/training-school-representation-mediated-multimodality/

[18] https://lr-coordination.eu/6thELRC6thELRCConference

[19] https://www.inlg2020.org/

[20] https://curso-linguaxe.pages.citius.usc.es/#programme

[21] https://gplsi.dlsi.ua.es/pln/node/553

[22] http://intranslation.mruni.eu/

[23] https://multi3generation.eu/workshops/eamt-2023/

[24] https://multi3generation.eu/outcomes/publications/

resources that they also deserve (low-resource languages). We understand that these shortcomings of AI and NLP in terms of interdisciplinarity and languages coverage led to trends and applications that present researchers and developers with the feeling of uncertainty of where and how to move next, especially with regard to human control of AI and generation of language without any form of discrimination. In the spirit of the interdisciplinarity of the COST Action Multi3Generation, we believe that more research is needed in order to create new and/or better resources and tools that capture human/expert knowledge required to build knowledge-based technologies (systems and products) and make sure that these systems are ethically developed, not developed with data that belongs to everyone being exploited by just a few. We envision the future of AI with knowledge-based linguistic methodologies being used and an increment in human resources from the human sciences, allowing linguists to share the drive in a process that has language as the object of these resources and tools. These human experts will capture human/expert knowledge required to build knowledge-based technologies (systems and products) and decide on the linguistic integrity of a language system, independently of the model that is under development. Our main concern is whether AI can/should be achievable without incorporating the science of language in all its branches, including psycholinguistics, social linguistics, among others.

AUTHOR CONTRIBUTIONS

Anabela Barreiro, Multi3Generation's chair, coordinated the article, organized the structure of the manuscript, wrote the Abstract and all Sections of the article, except the totality of Section II, and revised the article as a whole.

Elena Lloret, Multi3Generation's vice-chair, wrote part of Section II, commented on some sections of the article and revised the article as a whole.

Oleksii Turuta, Multi3Generation's working group 3 leader, wrote part of Section II, commented on some sections of the article and revised the article as a whole.

ACKNOWLEDGMENT

In this article, we acknowledge COST Action Multi3Generation (CA18231). Elena Lloret acknowledges also the following R&D projects: CORTEX (PID2021-123956OB-I00), funded by MCIN/AEI/10.13039/501100011033/ and by ERDF; "CLEAR.TEXT (TED2021-130707B-I00), funded by MCIN/AEI/10.13039/501100011033 and European Union NextGenerationEU/PRTR; and the project NL4DISMIS with grant reference CIPROM/2021/21, funded by the Generalitat Valenciana.

Position Papers of the 18th Conference on Computer Science and Intelligence Systems pp. 35–41
DOI: 10.15439/2023F6427
ISSN 2300-5963 ACSIS, Vol. 36

Assessing the Accuracy of Body Measurements through Regression Analysis

Janis Bicevskis
0000-0001-5298-9859
Faculty of Computing
University of Latvia
Email: Janis.Bicevskis@lu.lv

Edgars Diebelis
0000-0002-5950-9915
Faculty of Computing
University of Latvia
Email: Edgars.Diebelis@lu.lv

Zane Bicevska
0000-0002-5252-7336
Faculty of Computing
University of Latvia
Email: Zane.Bicevska@lu.lv

Ivo Oditis
0000-0003-2354-3780
Faculty of Computing
University of Latvia
Email: Ivo.Oditis@lu.lv

Girts Karnitis
0000-0003-2354-3780
Faculty of Computing
University of Latvia
Email: Girts.Karnitis@lu.lv

Oskars Ozols
0009-0006-0813-1808
DIVI Grupa Ltd
Email: Oskars.Ozols@ di.lv

***Abstract*—The digitalization of individual garment pattern construction presents challenges in accurately obtaining body measurements and constructing patterns tailored to specific individuals. This paper addresses the technological and conceptual aspects of transitioning from traditional, in-person tailoring to remote, digital pattern creation. It explores the need for algorithms that describe pattern construction operations in a computationally executable manner and the reliance on self-measurements by clients or their trusted individuals. The study focuses on evaluating the reliability of self-measurements and the potential errors introduced in the pattern construction process. The paper proposes the use of regression analysis to identify suspicious or erroneous measurement sets and assess their impact on the resulting garment shape. The study investigates the hypotheses regarding the identification of incorrect measurements through regression analysis and the application of publicly available artificial intelligence solutions. The findings contribute to enhancing the precision and reliability of digital individual garment pattern construction, facilitating remote creation and production processes.**

***Index Terms*—Quality control of graphical images, regression testing, regression analysis**

I. Introduction

THE digitalization of the individual garment pattern construction process presents technological and conceptual challenges. It requires the identification of algorithms that can describe the basic construction and modelling operations of patterns in a computationally executable manner. Additionally, significant changes need to be made to the garment pattern creation process and the techniques employed. Unlike in the widely used practice of pattern construction based on standard body measurements in the garment industry, individual garment construction is tailored to each person's specific body measurements. Traditionally, this process involves:

(1) taking body measurements performed by a professional tailor,
(2) constructing individual patterns for specific garment models,
(3) cutting the garment from fabric,
(4) fitting the garment through one or multiple iterations (including adjustments),
(5) finalizing the garment.

This process has been employed in tailoring for centuries but relies on close collaboration between the client and tailor, requiring them to be present in the same space and time.

With the advancement of digital technologies and the globalization of manufacturing, remote creation of individual products becomes increasingly relevant. In the field of garment sewing, this involves two aspects: (a) individual (perfect fit) garment construction in a digital environment and (b) individual garment production in specialized factories based on digital patterns. By transitioning the described process to a digital environment, pattern creation steps (Steps 1, 2, and 4) should be facilitated digitally, while the physical garment manufacturing steps (Steps 3 and 5) can be outsourced to service providers. The in-person meetings between the client and seamstress no longer take place, and the iterative refinement of the product is no longer possible. As a result, there are heightened requirements for both input data and precision in pattern construction.

The number of prepared patterns is enormous (over 175 different clothing models for more than 1000 clients), and their inspection requires significant resources. Automated pattern inspection is proposed, where the area and perimeter of each pattern element are calculated, which are closely related. In the research, the hypothesis is being confirmed that the ratio of perimeter to area is "invariant". This allows, firstly, to control the quality of newly entered data by utilizing previously accumulated models and customer indicators. Secondly, the ratio's numerical value can be used in

This work has been supported by University of Latvia projects: AAP2016/B032 “Innovative information technologies"

regression testing to ensure that changes in programs do not cause significant alterations in already accepted patterns.

This paper is structured as follows: problem statement (Section 2), related research (Section 3), experimental design and data set (Section 4), analysis of experiment (Section 5) and conclusion (Section 6).

II. PROBLEM STATEMENT

One of the significant challenges is evaluating the reliability of body measurements obtained in the digital environment according to the specific needs of the pattern construction method they use. In the digital environment, measurements must either be determined automatically (e.g., from photographs or using 3D scanners) or rely on self-measurements by the client (usually non-professionals). Automated body measurement determination has been extensively studied, but unfortunately, the results are unreliable in practice. Various factors contribute to this, such as the quality of the photographs, imprecise posing, the clothing worn during photography, limitations of image recognition algorithms, restrictions on the transmission of personal data, and others. This topic merits separate publications and will not be further discussed here.

The focus of this study is on measurements taken by the clients themselves or their trusted individuals. Several problems can be observed:

(1) different pattern construction methods require measurements to be taken in various ways,
(2) individuals without sewing experience may struggle to measure the body accurately, resulting in measurements that are too tight or loose, taken in incorrect locations, and so on,
(3) individuals cannot measure certain body dimensions themselves, thus relying on assistance from others, leading to stress and additional errors.

As a result, incorrect measurements (not corresponding to the specific body) may be obtained, resulting in garments that do not fit the individual, even if the pattern construction algorithm is flawless.

If suspicious sets of measurements or measurement's sets likely to have resulted from erroneous actions could be identified in an automated manner, it would be possible to reduce the risk of producing non-fitting (mismatched) patterns. One of the methods for identifying problematic measurement sets could involve establishing mandatory relationships between measurement definitions and the activation of control mechanisms at the time of measurement registration. Utilizing these relationships could filter out blatant errors, such as an impossible scenario where a woman's bust circumference is smaller than the under bust circumference. However, this approach does not aid in statistically identifying combinations of measurements that are highly unlikely due to the significant variability in human bodies. It is nearly impossible to find universal measurement relationships based solely on experience.

In this study, measurement relationships were analyzed using regression analysis and the capabilities of artificial intelligence on a collection of historically accumulated sets of body measurements. The following hypotheses were tested:

(Q1) Incorrectly entered individual measurements can be identified using regression analysis.

(Q2) Patterns constructed based on incorrectly entered measurements resulting in unusual garment shapes can be identified using regression analysis.

(Q3) Incorrectly entered individual measurements can be identified using publicly available artificial intelligence solutions.

III. RELATED RESEARCH

Accurate and reliable body measurement recognition is of paramount importance in diverse fields such as fashion, healthcare, ergonomics, and virtual reality. The ability to precisely capture and analyse body measurements plays a crucial role in personalized product design, fit optimization, and user experience enhancement. In recent years, significant advancements have been made in the field of body measurement recognition, leveraging cutting-edge technologies.

Traditional methods of obtaining body measurements often rely on manual measurement techniques performed by trained professionals. However, the advent of digital technologies has paved the way for alternative approaches that can enhance measurement accuracy, efficiency, and accessibility. Two prominent technological domains that have significantly impacted body measurement recognition are 3D scanning and image recognition.

A. Research on Image Recognition

There are several papers providing insights into the state-of-the-art techniques, algorithms, and challenges in the field of body measurement recognition from images [1] is a survey providing an overview of various techniques and approaches used for human body measurement estimation from images. [2] focuses on the application of image processing techniques for human body measurement and virtual try-on of clothing; it presents algorithms and methods for extracting body measurements accurately from images and simulating the try-on experience virtually.

[3] provides an overview of image processing techniques used for automatic human body measurement. It discusses various image analysis methods, feature extraction algorithms, and measurement estimation techniques employed in this field.

[4] explores the application of deep learning techniques for estimating human body measurements from images. It discusses the use of convolutional neural networks (CNNs) and other deep learning architectures to achieve accurate and robust measurement estimation.

[5] paper focuses on body measurement extraction and analysis techniques for apparel online retailing. It discusses the use of computer vision and image processing algorithms

to extract accurate body measurements from customer images and analyse them for personalized clothing recommendations.

B. Anthropometry and 3D Scanning

Anthropometry, the measurement of human body dimensions, plays a crucial role in diverse fields such as ergonomics, clothing design, healthcare, and biometrics. With advancements in technology, 3D body scanning has emerged as a powerful tool for capturing precise body measurements, offering a more comprehensive and accurate alternative to traditional measurement techniques. The "IEEE IC 3DBP" dataset [6] provides researchers with a valuable resource for comparative analysis and benchmarking of different anthropometric methods in 3D body scanning.

[7] is a study focusing on comparing different anthropometric measurement techniques based on 3D body scanning and [8], [9] explores the development of a population-specific anthropometric model based on 3D body scanning data. The findings highlight the importance of considering population-specific variations in anthropometric analysis for applications such as clothing design, ergonomics, and product development.

[10] investigates the application of machine learning algorithms in anthropometric analysis using 3D body scans. The research explores the use of machine learning models for automated feature extraction, body segmentation, and prediction of anthropometric measurements. The findings demonstrate the potential of machine learning techniques in improving the efficiency and accuracy of anthropometric analysis in large-scale datasets.

[11], [12], [13] - conduct a comparative analysis of different 3D body scanning technologies and sensor technologies". The study evaluates the performance, resolution, and accuracy of various scanning techniques, such as structured light scanning, laser scanning, and depth sensing.

[14], [15] aims to validate the accuracy and reliability of 3D body scanning measurements by comparing them with traditional anthropometric methods. The research investigates the agreement between measurements obtained from 3D body scans and manual measurements using clippers and tape measures. The findings contribute to establishing the validity and practicality of 3D body scanning as a reliable measurement technique.

IV. Experimental Design and Data Set

Within the project, more than 175 clothing pattern models have been developed. To ensure the precise fit of the intended clothing model for the customer, appropriate measurements of the customer's body are necessary. A total of 53 different body measurements (part of them see Table 1) have been identified for the developed clothing models, which enable the preparation of various clothing patterns (pants, skirts, jackets, dresses, coats, tops, etc.).

Table 1.
List of used measurements

Code	Name
Ag	Body height
Gkra	Main bust circumference
Ga1	Hip circumference
Ga2	Hip-thigh circumference
Csa	Thigh circumference
Pca	Under knee circumference
PlslL	Shoulder slope Right
Clg	Knee level
PlslKr	Shoulder slope Left
Bg	Length of trousers and maxi skirt
GdS	Hip diameter – side view
PrB	Trousers and skirt balance - Front
GdPr	Hip diameter – front view
MB	Trousers and skirt balance - Back
Ka	Neck circumference
Cstl	Groin arc
Etc.	53 in total

Each specific clothing model requires a certain set of measurements. For example, pants do not require measurements such as bust or neck circumference, but they do require hip and thigh measurements, among others. Not all 53 measurements are necessary for every clothing item. Each clothing pattern consists of multiple pieces.

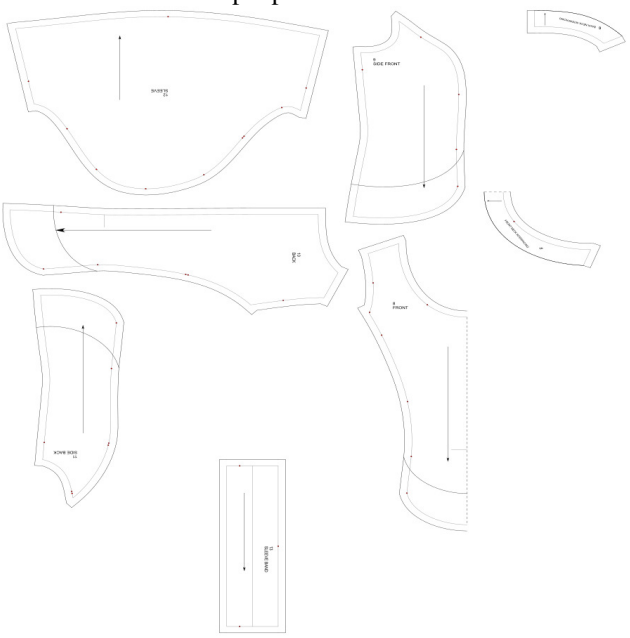


Fig. 1. Pieces of the Misses' and Women's Tops pattern

Within the scope of the study, the Misses' and Women's Tops clothing pattern was chosen, which consists of 8 pieces (see Figure 1) and utilizes 25 measurements: Ag, Gkra, PlslL, PlslKr, Ka, Plg, Trg, Pln, Ra1, Papla, Zkra, Va, Mg, Mpl, Prpl, Krg, Prgl, Krau, Zkrl, Ga1, Ga2, Sa, Elk, Pg, Ra2.

To ensure the accuracy of clothing fit or identify potential discrepancies, a comprehensive set of 274 customer measurements was used. In the following two tables (see Table 2), an example of measurements for 7 different customers is provided. The measurement GID is not included in the tables as it was not indicated for the sample customers and was not used in the data analysis of the study.

Table 1.

Example 1: Measurements of seven users (1)

	Aiga	Airita	Brandi	DANA	FT_Olivia	DINA	FT_DJ
Ag	174	173	150	170	175	164	164
ApmE	139	140		113		117	
ApmS	127	133		107		102.5	
Bg	104.5	103.5	91.44	103	99.6	102.5	104

The empty cells in Table 2 indicate that a specific measurement is not assigned to a particular customer.

To evaluate the accuracy or discrepancies in the patterns, the article suggests comparing customer measurements with two pattern parameters: the length of the pattern contour line (perimeter) and the area in pixels. An example of calculating the line length and area is shown in the following figure (Fig. 2)

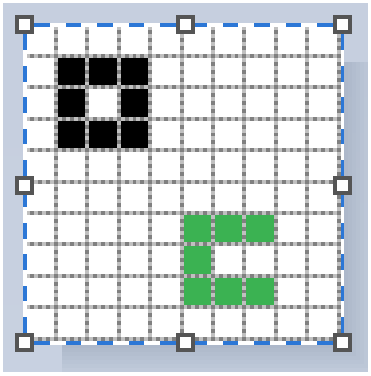


Fig. 2. Example of perimeter and area calculation

In the image (Fig. 2), two examples of pattern components are visible. The black component has an area of 9 pixels and a perimeter of 12 pixels, while the green component has an area of 7 pixels and a perimeter of 16 pixels. Using the aforementioned algorithm for calculating the area and perimeter of pattern components, the area and perimeter were calculated for the eight components of the mentioned pattern for a total of 274 customers in the study. Table 4 provides the calculated areas and perimeters in pixels for the eight pattern components for three selected customers.

Table 2.

Area and perimeter in pixels of the pattern pieces

Customer		Piece 1	Piece 2	Piece 3	Piece 4	Piece 5	Piece 6	Piece 7	Piece 8
Aiga	Area (px)	1129874	1392457	1781631	1065098	2934855	805799	218969	151424
	Perimeter (px)	17630	5906	7468	5516	8430	4180	3172	2268
Airita	Area (px)	1156991	1497732	1700162	1054941	2838373	821524	208735	155045
	Perimeter (px)	17628	6036	7132	5422	8402	4246	3058	2324
Brandi	Area (px)	7381110	9291294	1344574	7846754	2038087	714123	194394	158019
	Perimeter (px)	15418	4962	6596	4648	7446	3824	2882	2312

V. ANALYSIS

A. Regression Analysis of Pattern Data

The examined dataset includes multiple measurements, some of which are correlated with each other. It is evident that the lengths of a person's right and left arms or legs are correlated. This provides an opportunity to calculate potential correlations and identify outliers from the available data. The dataset consists of measurements for 274 individuals, but all measurements are available for 100 individuals.

Correlation analysis was performed for these individuals, and Pearson correlation was calculated for each pair of measurements. The correlation was above 0.9 for 55 pairs of measurements. Scatterplots were created for these pairs, showing both the mutual dependencies of the measurements and specific outliers.

For example, in Fig. 1, there is a strong correlation between parameters Csa and Pca, but there is an outlier with

coordinates (61, 49). In Fig. 2, it can be observed that there is a pronounced correlation between Ga1 and Ga2, but there is an outlier with coordinates (88.5, 70). Such outliers serve as a serious warning that there may be errors in the entered data. Though it cannot be stated with absolute certainty, it can serve as a basis for the system to verify the entered data during data entry to ensure their accuracy.

For each piece, the Pearson correlation was calculated between the piece's area and its perimeter. Although the area should ideally be quadratically dependent on the linear dimensions of the piece, all pieces of the examined model showed a high level of linear correlation between the area and perimeter of the piece. This was most pronounced for piece 5 (Fig. 5). On the other hand, scatterplot for Piece 3 Perimeter vs Area (Fig. 6) demonstrates some nonlinearity. When examining the scatterplot of each piece, where one dimension represents the piece's area and the other dimension represents the piece's perimeter, outliers were discovered in several measurements of pieces (piece 1, piece 4) (Fig. 7, Fig. 8). The outlier for piece 1 with coordinates (5002, 678011) belongs to the same individual who had a suspicious ratio of measurements Ga1 and Ga2. It is possible that the atypical measurements resulted in the creation of a piece with an atypical shape. On the other hand, the outlier of piece 4 with coordinates (10270, 2703263) belongs to piece for different individual (MF-Laura Š), that could show the problem with generated piece.

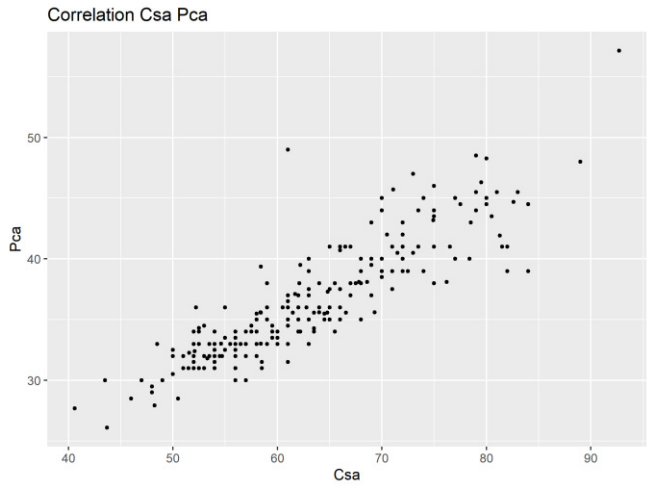


Fig. 3. Correlation Csa === Pca

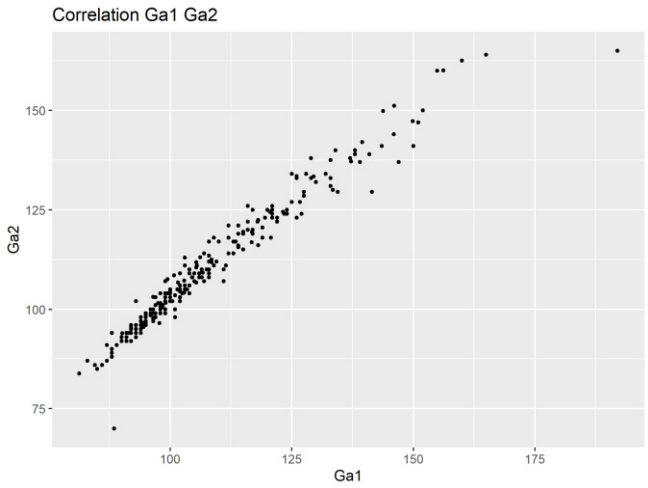


Fig. 4. Correlation Ga1 === Ga2

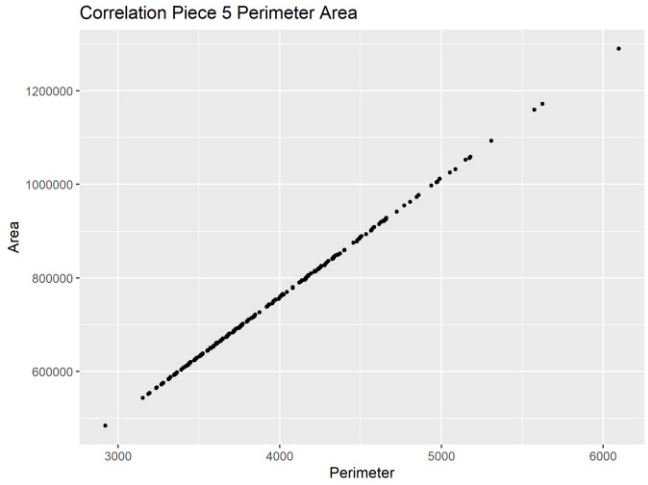


Fig. 5. Correlation Piece 5 Perimeter === Area

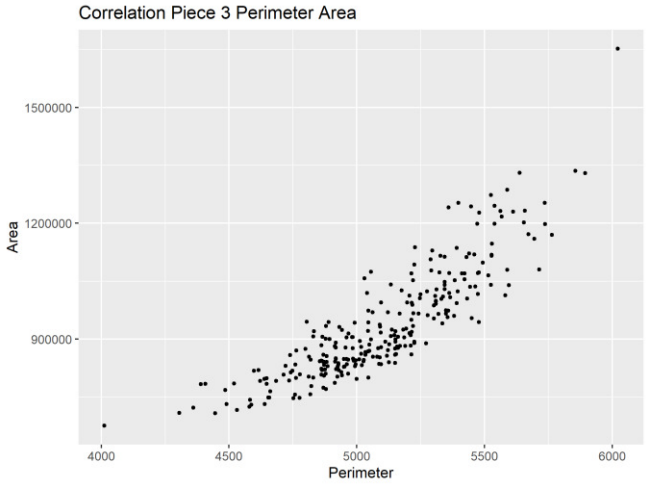


Fig. 6. Correlation Piece 3 Perimeter === Area

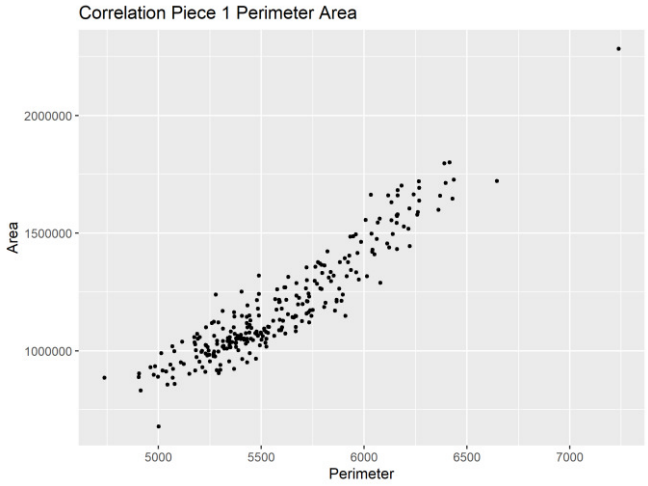


Fig. 7. Correlation Piece 1 Perimeter === Area

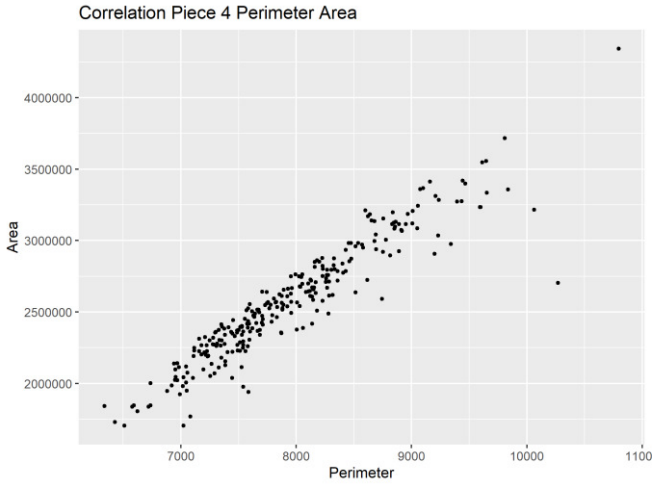


Fig. 8. Correlation Piece 4 Perimeter === Area

B. Analysis of Data Using ChatGPT

In recent times, there has been significant interest in utilizing artificial intelligence for identifying and analyzing various characteristic patterns. In our research, we also opted to employ artificial intelligence to analyze our data. This approach could allow us to gain a broader and more in-depth insight into the data structure and uncover potential correlations and characteristic trends that could be relevant to our research domain. For our study, we employed the freely available ChatGPT [16] tool provided by OpenAI [17]. Our objective was to analyze a dataset of human body measurements. However, the tool did not identify any outliers. Instead, it highlighted specific values that deviated from the overall dataset.

For example:

"Prof 1": The body height of 164cm was relatively high compared to majority of the dataset, which primarily ranged between 156cm and 173.5cm.

"Prof 2": The bust volume of 38 stood out as notably higher than the other values, which ranged between 31 and 36.7.

It's important to note that the tool did not provide a more comprehensive analysis of the dataset.

To improve results and enhance analysis, datasets for ChatGPT should be preformatted to provide additional context. Even when explicitly entering incorrect values, such as a height of 264cm, the tool only recognized the issue after performing a specific task to validate all body heights. After multiple iterations and attempts, we settled on the following prompt: "Analyze the quality of human body measurement data for sewing, identifying unusual values. The dataset contains measurements in centimeters, with column headers on the first line." However, it should be noted that the tool yielded different results with each run.

To utilize ChatGPT for data analysis, we would need to collect additional personal characteristic data such as weight, gender, geographic affiliation, and other data, which are currently unavailable to us.

VI. CONCLUSION

A method is proposed for automated quality control of individually tailored garment patterns. The essence of the approach is as follows: first, the designers create clothing models and present them to clients. From the various models, the client chooses the most suitable one and places an order for its production. This, in turn, requires the client to provide their body measurements, which can be significant in number, even up to 53. The designer then programs a pattern generation algorithm that considers the client's measurements and is capable to generate high-quality garment patterns. Individual garments are then sewn from these patterns, providing the client with a more suitable fit compared to mass-produced options.

Unfortunately, the number of prepared patterns is enormous (over 175 different clothing models for more than 1000 clients), and their inspection requires significant resources. Automated pattern inspection is proposed, where the area and perimeter of each pattern element are calculated, which are closely related (high correlation coefficient). If the correlation coefficient is below a critical threshold, individual pattern inspection is required. This method can also detect inaccuracies in the input of client measurements.

REFERENCES

[1] Ruiz N., Bueno M.B., Bolkart T., Arora, Lin M., Romero J., Bala R. Human body measurement estimation with adversarial augmentation. International Conference on 3D Vision, 2022 https://www.amazon.science/publications/human-body-measurement-estimation-with-adversarial-augmentation (Accessed 23.05.2023).

[2] Roknabadi A.D., Latifi M., Saharkhiz S., Aboltakhty H. Human body measurement system in clothing using image processing. World Applied Sciences Journal 19(1):112-119 DOI: 10.5829/idosi.wasj.2012.19.01.1306, January 2012.

[3] Liu X., Wu Y., Wu H. Machine Learning Enabled 3D Body Measurement Estimation Using Hybrid Feature Selection and Bayesian Search. Appl. Sci. 2022, 12(14), 7253; https://doi.org/10.3390/app12147253.

[4] Ashmawi S., Alharbi M., Almaghrabi A., Alhothali A. Fitme: Body Measurement Estimations using Machine Learning Method. Procedia Computer Science. Volume 163, Pages 209-217, 2019. https://doi.org/10.1016/j.procs.2019.12.102.

[5] Bye E., Labat K.L., Delong M. R. Analysis of Body Measurement Systems for Apparel. Clothing and Textiles Research Journal 24(2):66-79, March 2006 DOI: 10.1177/0887302X0602400202
[6] https://ieee-dataport.org/open-access/dataset-ieee-ic-3dbp-comparative-analysis-anthropometric-methods.
[7] Bartol K., Bojanić D., Petković T., Pribanic T. A Review of Body Measurement Using 3D Scanning, IEEE Access, DOI: 10.1109/ACCESS.2021.3076595, 2021.
[8] Lu J., Wang M.J. Automated anthropometric data collection using 3D whole body scanners, DBPL, Expert Systems with Applications 35(1-2):407-414, July 2008. DOI: 10.1016/j.eswa.2007.07.008
[9] Kuribayashi M., Nakai K., Funabiki N. Image-Based Virtual Try-on System With Clothing-Size Adjustment. DOI: 10.48550/arXiv.2302.14197, 2023.
[10] Pleuss J.D., Talty K., Morse S., Kuiper P., Scioletti M., Heymsfield S.B., Thomas D.M. A machine learning approach relating 3D body scans to body composition in humans. Eur J Clin Nutr. 2019 Feb; 73(2): 200–208, published online 2018 Oct 12. doi: 10.1038/s41430-018-0337-1
[11] Bartol, K., Bojanić, D., Petković, T., Peharec, S., Pribanić, T. Linear Regression vs. Deep Learning:A Simple Yet Effective Baseline for Human Body Measurement. Sensors, 22, 1885, 2022. https://doi.org/10.3390/s22051885
[12] Kus A., Unver E., Taylor A. A Comparative Study of 3D Scanning in Engineering, Product and Transport Design and Fashion Design Education. Computer Applications in Engineering Education 17(3):263 – 271, September 2009DOI: 10.1002/cae.20213
[13] Seifert, E., Griffin, L. Comparison and Validation of Traditional and 3D Scanning Anthropometric Methods to Measure the Hand. Paper presented at 11th Int. Conference and Exhibition on 3D Body Scanning and Processing Technologies. https://doi.org/10.15221/20.41 , 2020.
[14] Skorvankova, D., Riečický, A., Madaras, M. Automatic Estimation of Anthropometric Human Body Measurements. 17th International Conference on Computer Vision Theory and Applications. (2021) DOI: 10.5220/0010878100003124, https://www.scitepress.org/PublishedPapers/2022/108781/108781.pdf
[15] Rumbo-Rodríguez L, Sánchez-SanSegundo M, Ferrer-Cascales R, García-D'Urso N, Hurtado-Sánchez JA, Zaragoza-Martí A. Comparison of Body Scanner and Manual Anthropometric Measurements of Body Shape: A Systematic Review. Int J Environ Res Public Health. 2021 Jun 8;18(12):6213. doi: 10.3390/ijerph18126213
[16] ChatGPT May/, 12 Version, https://help.openai.com/en/articles/6825453-chatgpt-release-notes
[17] OpenAI. https://chat.openai.com

Position Papers of the 18th Conference on Computer Science and Intelligence Systems pp. 43–50
DOI: 10.15439/2023F5494
ISSN 2300-5963 ACSIS, Vol. 36

Towards Community-Driven Generative AI

Rustem Dautov
0000-0002-0260-6343
SINTEF Digital
Forskningsveien 1
0373 Oslo, Norway
rustem.dautov@sintef.no

Erik Johannes Husom
0000-0002-9325-1604
SINTEF Digital
Forskningsveien 1
0373 Oslo, Norway
erik.johannes.husom@sintef.no

Sagar Sen
0000-0002-5784-7355
SINTEF Digital
Forskningsveien 1
0373 Oslo, Norway
sagar.sen@sintef.no

Hui Song
0000-0002-9748-8086
SINTEF Digital
Forskningsveien 1
0373 Oslo, Norway
hui.song@sintef.no

***Abstract*—While the emerging market of Generative Artificial Intelligence (AI) is increasingly dominated and controlled by the Tech Giants, there is also a growing interest in open-source AI code and models from smaller companies, research organisations and individual users. They often have valuable data that could be used for training, but their computing resources are limited, while data privacy concerns prevent them from sharing this data for public training. A possible solution to overcome these two issues is to utilise the crowd-souring principles and apply federated learning techniques to build a distributed privacy-preserving architecture for training Generative AI. This paper discusses how these two key enablers, together with some other emerging technologies, can be effectively combined to build a community-driven Generative AI ecosystem, allowing even small actors to participate in the training of Generative AI models by securely contributing their training data. The paper also discusses related non-technical issues, such as the role of the community and intellectual property rights, and outlines further research directions associated with AI moderation.**

***Index Terms*—Generative AI, Federated Learning, Crowd-Sourcing, Community, Conceptual Architecture, AI Moderation.**

I. Introduction and Motivation

GENERATIVE Artificial Intelligence (AI) refers to AI models that can generate original content, such as text, images, and music. Unlike traditional AI models that are trained to recognise and classify existing data, Generative AI models learn to generate new data by analysing patterns and structures in large datasets. ChatGPT, developed by OpenAI and sponsored by Microsoft, is admittedly the most prominent example of Generative AI, while similar proprietary services are also developed by the other Big Tech companies.[1]

At the same time, there is a growing interest in open-source AI from smaller companies, research organisations and individual users. It is, admittedly, not feasible for such small players to compete with the Tech Giants individually, but what if they could join forces to collectively challenge the establishing monopoly and lead the development of Generative AI using community-driven democratic principles?

[1]Please note that the main focus of this paper is on large language models (LLMs) as the most prominent and representative example of Generative AI, albeit the discussed concepts are applicable to a certain extent to other types of AI-generated content, such as imagery and sound.

A. Motivating Example: Assisted Code Generation

The fact that these leading tools are proprietarily owned or backed up by big corporations has major implications for their usage and development. One of the biggest concerns is the presence of *bias*, which not only naturally rises from the data used to train the AI and the training algorithm but is also artificially introduced in favour of the corporations' commercial or political interests. This 'intentional' bias can influence consumers who rely on Generative AI tools to make decisions. Another source of bias is filtering, which in theory is supposed to ensure that the generated content meets certain criteria and is appropriate for its intended use. In practice, however, the companies tend to play it safe and apply excessive filtering just to protect themselves from possible ethical scandals. While this is understandable, enforcing such filtering-based moderation may blur important aspects of reality. For example, an AI tool that filters out all mentions of a particular controversial topic may not accurately represent the diversity of opinions.

Another source of bias is that these tools are usually trained only on publicly available data scraped from the Internet, and thus do not account for more specific and nuanced information that is only accessible to private users. For example, if an AI model is trained on public code repositories, it will not incorporate the valuable information exchanged in private corporate networks (*e.g.*, repositories, chats, issue trackers), although these are often considered a more trusted source of professional knowledge than semi-professional answers and informal discussions on Stack Overflow or Reddit. More specifically, in the realm of programming and code generation, these existing models trained on public data sources often overlook a wealth of insights and professional exchanges found in private corporate networks. These networks contain note just code examples, but also specialised practices and innovative solutions, serving as valuable repositories of programming knowledge. Yet, such smaller entities with their unique knowledge are left out from contributing due to privacy and security concerns, and their valuable information is thereby excluded from training. Taken together, these limitations can have significant implications for the accuracy and fairness of the AI-generated output.

©2023, PTI

Thematic track: Distributed Edge AI – Risks and Challenges

B. Paper Contribution and Structure

With this paper we make a first step towards democratising and de-monopolising this emerging market by designing a community-driven Generative AI architecture. The proposed conceptual architecture relies on several existing technologies, which collectively represent a promising toolkit for building a whole open ecosystem for Generative AI. Some key features of the envisioned solution are the ability to preserve data privacy, unbiased and fair model training, decentralised operation, and transparent content moderation, among others. We claim that the emerging field of Generative AI should not be monopolised by the few Tech Giants, but rather collaboratively developed and moderated by an open community of multiple stakeholders, each providing their own perspective on this challenging, yet exciting technology. In explaining the envisioned approach, we also draw parallels with the core elements of democracy to better communicate the proposed concepts and ideas.

The main contribution of this paper is a conceptual architecture of a community-driven ecosystem for Generative AI. The description of this architecture is organised as follows. Section II presents the main technologies underpinning the design of the proposed architecture and describes their roles and benefits. Section III draws parallels with similar relevant projects and critically discusses assumptions and some further research considerations. Section IV summarises the paper with some concluding remarks.

II. Technological Building Blocks

We now present the envisioned conceptual architecture by describing its individual 'building blocks', as depicted in Fig. 1. The architecture can be seen as a vertical stack of technologies, which, we believe, provide a viable foundation for building a community-driven Generative AI ecosystem. The individual layers of the proposed stacked architecture build upon one another, each providing technological foundation for building the next layer. This layered structure is explained in the following subsections starting from the very bottom layer of hardware infrastructures.

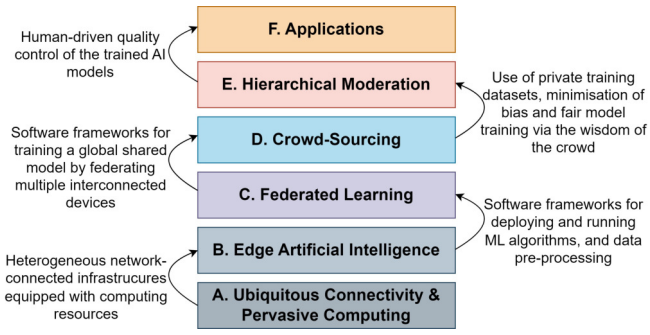


Fig. 1. Main elements of a community-driven Generative AI ecosystem.

A. Ubiquitous connectivity and pervasive computing

Recent technological advances have paved the way for *ubiquitous connectivity* [1] and *pervasive computing* [2] – the two concepts which revolve around the idea of seamless and pervasive access to computing resources and services. With the pervasive availability of network connections, devices and systems are enabled to be seamlessly connected to the Internet or other communication networks. It emphasises the widespread access to high-speed Internet, wireless networks, and advanced communication technologies. The goal of ubiquitous connectivity is to ensure that people and devices can communicate and access information from anywhere, at any time. This connectivity enables the exchange of data, collaboration, and interaction among various devices and systems.

At the same time, pervasive computing extends the concept of ubiquitous connectivity by focusing on the integration of computing capabilities into everyday objects and environments. It involves embedding intelligence into a wide range of personal devices and human-centred spaces, such as smartphones, 'wearables', household appliances, vehicles, buildings, *etc.*. The goal is to create an environment where computing and information processing become seamlessly integrated into people's daily lives, without requiring explicit user intervention. Together, these technological trends support the growth of emerging technologies such as the Internet of Things (IoT), smart cities, autonomous vehicles, real-time analytics, and other applications requiring low latency, high reliability, and efficient use of network resources [3].

Foundation for the next layer: Ubiquitous connectivity and pervasive computing provide a distributed infrastructure of heterogeneous network-connected devices equipped with computing resources.

B. Edge Artificial Intelligence

The described advances in networking and computing capabilities of field-deployed devices underpin another relevant concept – *edge computing*, which is a decentralised computing paradigm that brings data processing and computation closer to the data source or the 'edge' of the network, instead of relying solely on centralised cloud servers [4]. In edge computing, data processing and analytics are performed at or near the device/sensor level instead of sending all data to a remote data centre for processing. The advantages of edge computing include:

- *Reduced latency*: By processing data locally at the edge, response times can be significantly improved, enabling near-real-time applications.
- *Bandwidth optimisation*: Edge computing reduces the amount of data that needs to be transmitted to the cloud or data centre, thus optimising network bandwidth usage and reducing costs.
- *Enhanced privacy and security*: Local data processing at the edge can help protect sensitive information by reducing the exposure of data in transit and allowing for localised security measures.
- *Offline functionality*: Edge devices can continue to operate and provide services even when connectivity to the cloud is disrupted, ensuring uninterrupted functionality.

Data processing at the edge can range from simple data pre-processing operations to rather advanced AI analytics using Machine Learning (ML). The latter, commonly known as *Edge AI*, refers to the deployment of AI algorithms and models directly on edge devices, such as smartphones, IoT devices, edge servers, and other similar computing nodes [5]. It brings AI capabilities and decision-making closer to the data source, minimising the need for data transmission to centralised cloud servers. The key idea behind Edge AI is to run complex ML-driven data analytics locally at the edge device itself, without relying on continuous cloud connectivity or sending data to remote data centres. This enables near-real-time inference, reduces latency, saves bandwidth, enhances privacy, and enables offline functionality even in the absence of the Internet connection. All these features are especially important to the healthcare domain where physiological data collected by wearable or portable medical devices are processed either directly on those devices or on a smartphone acting as a wireless gateway [6], [7]. Similarly, the data privacy and network bandwidth constraints are usually critical aspects in various image and video recognition scenarios involving CCTV cameras [8], [9].

Foundation for the next layer: Edge AI provides software frameworks for deploying and running ML algorithms, as well as data pre-processing on top of heterogeneous, potentially resource-constrained edge hardware infrastructures.

C. Federated Learning

A natural next step in the Edge AI development was not only to deploy pre-trained AI models and run inference, but also to train the models at the edge. While individual edge devices are still constrained in their computing capabilities to perform heavy-weight model training, the promising solution was to combine multiple devices into an aggregated pool of computing resources and then orchestrate the iterative model training process. This ML approach, known as *federated learning*, enables training models on decentralised data without the need to transfer raw data to a central server [10]. In federated learning, the training process takes place directly on edge devices, such as smartphones, IoT devices, or local servers, where the data is generated and stored. The main idea is to bring the model training to the data rather than to move the data to a central location.

Some prominent federated learning frameworks actively developed and used by the community include Flower,[2] Tensorflow Federated,[3] and OpenFL.[4] Federated learning has applications in various domains, including healthcare, finance, smart devices, and more. It allows for collaborative learning while maintaining data privacy, making it a promising approach for training models on sensitive or distributed data sources. In the context of Generative AI, federated learning can be applied to train LLMs in a distributed and privacy-preserving manner. Instead of centralising the training data on a single server, training can be performed directly on edge devices or local servers where the data resides. The main benefits of federated learning applied to Generative AI include:

- *Privacy*: Federated learning preserves data privacy since the raw data remains on the edge devices and is not directly shared with the central server. This is particularly important when dealing with sensitive user data.
- *Data localisation*: Federated learning enables training on data that is distributed across multiple devices or locations, allowing for localised training and personalised models while avoiding data silos.
- *Efficiency*: Training LLMs can generate a massive amount of data, making communication between devices and the central server resource-intensive. By training models on edge devices, federated learning reduces the need for data transmission over the network, saving bandwidth and lowering communication costs.
- *Distributed computing power*: By promoting decentralised ML, federated learning enables local nodes to participate in the training process. This can improve responsiveness, reduce latency, and enhance autonomy. As a result, leveraging the computing power of multiple devices or servers enables faster training of LLMs by parallelising the training process. Such pooled computational and storage resources federated across a sufficiently large set of participating nodes can even compete with the infrastructural computing resources of the Tech Giants.

To avoid a potential bottleneck and a single point of failure, the community has also proposed so-called *gossip learning* [11], [12] inspired by the gossiping behaviour observed in social networks, which involves the exchange of information and model updates among participating nodes, rather than with one central node. Gossip algorithms are distributed protocols used for information dissemination and aggregation in decentralised systems [13]. In a gossip algorithm, information spreads through the network by means of local peer-to-peer interactions. The algorithm operates in rounds or iterations, and in each round, a node (or a subset of nodes) selects one or more neighbours to exchange information with. Over time, the information spreads across the network as nodes continue to interact and share information with their neighbours. In addition to the default benefits of federated learning, gossip-based extensions provide the following benefits:

- *Scalability*: Gossip learning can scale well with large networks, as each node only needs to communicate with a small number of other nodes at each iteration.
- *Fault tolerance*: Gossip learning is resilient to node failures or network partitions. Even if some nodes become unavailable, the information can still spread through the network via other nodes.
- *Adaptability to dynamic environments*: Gossip learning can adapt to changes in the network, such as nodes joining or leaving dynamically, allowing for continuous learning in evolving environments.

Gossip algorithms provide a decentralised and scalable

[2]https://flower.dev/
[3]https://www.tensorflow.org/federated/
[4]https://github.com/securefederatedai/openfl/

approach for communication in federated learning, allowing devices to collectively learn from each other while preserving data privacy. They can also handle communication failures, device churn, and heterogeneity in device capabilities. Various gossip algorithms, such as random pairwise gossip, ring-based gossip, or hierarchical gossip [14], can be employed depending on the specific requirements and characteristics of the federated learning scenario. The main steps of a federated learning setup, enhanced with gossip algorithms, are the following (depicted in Fig. 2):

1) *Initialisation*: Each participating device initialises its local model with an initial set of parameters.
2) *Local training*: Each device trains its local model using own data, following a predefined training process. This can involve multiple training iterations (*i.e.*, epochs).
3) *Communication and model exchange*: A subset of devices is selected to participate in the communication process. The selection can be random or based on certain criteria, such as device proximity or resource availability. During the communication round, selected devices exchange information, which can include sharing model parameters, gradients, or other relevant updates.
4) *Update aggregation*: Next, each device updates its local model by aggregating the received information from other devices. The aggregation process can vary and may include techniques like averaging, weighted averaging, or more sophisticated aggregation strategies [10].
5) *Repeat*: Steps 2-4 are repeated for multiple communication rounds or until convergence criteria are met. The goal is to iteratively refine the local models and improve the global model without sharing raw data.

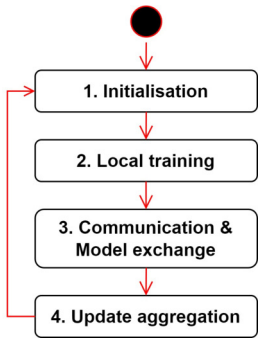


Fig. 2. Federated learning workflow based on crowd-sourced training data.

Foundation for the next layer: Federated learning provides software frameworks for training a global model using private datasets from distributed interconnected devices.

D. Crowd-sourcing

Federated learning makes it technically possible for distributed clients to participate in a collaborative model training process. At the same time, it enables them to safely contribute their local, potentially sensitive datasets for training. In the context of Generative AI, these could be, for example, some technical documentation within an internal corporate network or photo images stored on a personal smartphone. The involvement of human users or organisations in a federated setup is strongly related to the concept of *crowd-sourcing* [15], which can be used for training LLMs by harnessing the collective efforts and knowledge of a diverse group of contributors.

In a broad sense, crowd-sourcing is a process of obtaining services, ideas, or content by soliciting contributions from a large group of people. It involves breaking down a global task into small, discrete parts and distributing those parts among participants, who then upon completion of their local parts, contribute the outputs to achieve the global task. In the context of LLM training, crowd-sourcing would assume the participation of a large and diverse group of online users contributing with their collected private datasets to train a global AI model.

Incentive mechanisms play a crucial role in crowd-sourced federated learning to encourage participation and cooperation among the participating users and organisations [16]. These mechanisms aim to align the interests of the participants with the overall objectives of the federated learning process. Some common incentive mechanisms applicable in this context can be based on monetary, reputation or resource rewards given to the crowd-sourcing participants. Admittedly, designing effective incentive mechanisms is a challenging task, as it requires balancing the objectives of individual participants with the global goals of the Generative AI system, at the same time ensuring fairness, privacy, and end-to-end security.

Another important aspect of crowd-sourcing, provided a sufficiently large number of participating actors, is the *diversity* – *i.e.*, crowd-sourcing allows for the inclusion of diverse perspectives and linguistic variations in the training data, enhancing the language model's ability to handle different languages, cultural contexts, competences, beliefs, etc. By leveraging the collective intelligence and efforts of a diverse group of contributors, crowd-sourcing enables the creation of more robust and inclusive models. Taken together, the sufficiently large number of participants and their diversity in a federated crowd-sourcing setup will enable the so-called *wisdom of the crowd*. The wisdom of the crowd is a concept that suggests that a group of individuals collectively can make better decisions or provide more accurate answers than an individual expert [17]. It is based on the idea that aggregating diverse opinions and knowledge from a large group leads to a more reliable and accurate outcome. The wisdom of the crowd relies on the inclusion of diverse independent viewpoints, opinions, and knowledge. This way, when individuals with different backgrounds and perspectives contribute their insights, it reduces biases and brings a wider range of information to the decision-making process. It is assumed that errors or biases present in individual opinions (*i.e.*, the minority) are cancelled out or outweighed by the collective judgement of the majority. The wisdom of the crowd is a crucial milestone in collective training of unbiased and fair Generative AI models.

Foundation for the next layer: Crowd-sourcing allows using private datasets for training going beyond the publicly

available data on the Internet, minimises bias and achieves fair model training results via the wisdom of the crowd.

E. Hierarchical Moderation

A critical aspect of Generative AI is its moderation. AI moderation refers to the process of regulating and controlling the behaviour of AI systems to ensure they operate in a responsible and ethical manner [18]. In an ideal scenario, efficiently implementing crowd-sourcing and achieving the described wisdom of the crowd will assume inherent self-moderation, where the diversity and the large number of participants will ensure that biases and errors of individual training inputs are balanced out by the strengths of others. This can be compared to the majority and plurality rules in democracy – the principles of taking most popular group decisions, where all expressed opinions are treated fairly by giving each an equal weight.

Humanity, however, knows many examples when the majority was wrong. Therefore, a more realistic scenario is the introduction of an additional moderation layer, which will rely on the democratic power separation principles, such that no single authority has the power to evaluate the accuracy of the data or the model (as it happens now with the mainstream Generative AI tools). *Hierarchical moderation* (or hierarchical governance) is a model of content moderation and decision-making that is structured in a hierarchical manner to ensure the quality and reliability of its articles [18]. A notable reference in this context is Wikipedia, which relies on a distributed hierarchy of community-nominated and elected editors to ensure the correctness and fairness of user-generated content. Implementing a similar automated moderation system for crowd-sourced LLMs would rely on training advanced algorithms to detect and remove harmful or offensive content. This could also involve creating separate models that are trained on inappropriate content, allowing to identify similar content and flag it for review. It is important to note that the moderation framework should respect the democratic principles and operate in a collaborative and consensus-driven manner. The community's input and involvement are key to maintaining the integrity and quality of the contents. The hierarchical structure, policies, and roles need to be designed to provide a framework for decision-making and to address the described content moderation and quality control issues.

Noteworthy, maintaining oversight in community-driven AI should not mean reinforcing entrenched biases or imposing a single view of the truth. Rather, it should nurture an environment that enables AI to continually learn, unlearn, and relearn in harmony with the evolving human insights. Essentially, the moderation process should mirror the fluid nature of understanding – a perpetual journey that encourages humility, values diverse perspectives, and nourishes a communal spirit. The focus should be on shared enlightenment and empathy, fostering an AI that augments common human experiences rather than homogenising or belittling them. This oversight includes intensive testing of updates via a validation set, real-time A/B testing, fairness, and bias evaluations, as well as adversarial testing. These tests should be performed by a diverse group of participants using their local validation sets. The results from these tests need to be collated to create a global metric that informs whether new knowledge introduced into the community-driven AI requires modification or complete exclusion.

Foundation for the next layer: Moderation enables producing fair and accurate language models with minimum bias or misinformation. This way, the models can be further safely used in various applications.

F. Applications

Finally, with the rest of the elements in place, it will be eventually possible to build the Generative AI software tools based on the collaboratively-trained models. The scope of such tools is not expected to differ from the currently being in use. While new applications and use cases continue to emerge on a daily basis, LLMs are already playing a key role in the following scenarios: various chat bots, virtual assistants and recommendation systems, content comprehension, generation and moderation, sentiment analysis and opinion mining, to name a few. It is expected that the unbiased and transparent nature of the community-driven Generative AI tools will create fair competition with the proprietary commercial tools, thus also facilitating the increased quality of the available products.

Coming back to the motivating example, the proposed community-driven Generative AI could provide an effective solution to build a more accurate and intelligent assisted coding functionality. Leveraging both crowd-sourcing and federated learning, organisations and individuals can contribute their unique knowledge to the AI model, including proprietary coding methodologies that may be less common in public and open-source code. This non-public data could come from private repositories, issue trackers and enterprise source code management systems. Federated learning ensures privacy of data by only sharing the updated model parameters, thereby preserving the privacy of proprietary and sensitive information. Simultaneously, the model becomes enriched with a diverse range of programming knowledge, extending beyond what public repositories can offer.

The practical implications of such an approach in the field of programming could be transformative. Potential benefits include more enhanced problem-solving capabilities, a richer understanding of proprietary coding practices and improved knowledge of lesser used programming languages. This community-driven model could democratise AI within the programming sphere, allowing even smaller contributors to play a role in AI development. Careful considerations of intellectual propoerty rights and the fostering of a robust community to drive this process forward are necessary for the successful implementation of this model. As such, the proposed community-driven Generative AI provides a promising alternative that not only addresses current limitations but also promotes fair competition with proprietary commercial tools.

III. Assumptions and Further Considerations

In this section, we discuss our concerns of non-technical nature, which should also be taken into account, as well research considerations going beyond the scope of this paper.

A. Role of the Community

Combined, the described technologies represent a powerful toolkit for building a community-driven Generative AI ecosystem, which can challenge the establishing plutocracy of big corporations. We have already seen examples of similar large-scale collaborative projects in the past. The most notable example is NASA's SETI@home project [19], which connected more than 5 million users contributing their private computing resourcing. A more recent example using federated learning is the MELLODY project,[5] which connected several medical research institutions into a federated learning network used for drug discovery in the pharmaceutical industry.

Also, the proposed vision shares many similarities with various open-source software foundations, such as Linux, Apache, and Eclipse. All these organisations are community-driven and follow the principles of transparent and open communication and code distribution. Therefore, an important assumption of this proposed vision is the active involvement of the community in establishing and further developing such an open ecosystem for Generative AI. As we already argued, even in the presence of crowd-sourcing and decentralised gossip learning, there is still a need for aggregating the model, developing the training algorithms, as well as the moderation – all these activities cannot (and should not) be fully decentralised.

B. Intellectual Property Rights

Intellectual property rights (IPR) is another important consideration in the context of envisioned architecture, as it involves collaboration and sharing of information among multiple parties, followed by generation of new creative content and its eventual consumption by end users. While there is still many open questions, the key considerations related to IPR can be summarised as follows:

- *Data and model ownership*: In a federated learning setup, the participants typically retain ownership of their data. This means that the data used for training the models remains under the control and ownership of the participants. At the same time, the resulting global model itself may be subject to IPR. The ownership of the model can vary depending on the agreements and arrangements between the parties involved. It is the responsibility of the community to establish clear guidelines and agreements regarding the ownership and use of the trained models.
- *Copyright*: Copyright law protects original creative works, such as text, images, music, or videos, from unauthorised copying, distribution, or use. In the case of Generative AI, questions arise regarding the ownership of AI-generated content. Typically, the copyright ownership is attributed to the human creator or the owner of the AI system, as they provide the input and training data for the AI model. In the proposed architecture, however, there is no such single actor. Therefore, it is again up to the community to decide and agree on the applicable copyright ownership policies.
- *Derivative works*: Generative AI models can be used to create derivative works based on existing copyrighted material (*i.e.*, private data crowd-sourced for federated learning). The legal implications associated with the creation and use of such derivative works need to be further explored, as they may require permission or licensing from the original copyright holder.

The application of IP laws to generative AI raises complex and evolving legal questions. Different jurisdictions may have different interpretations and regulations regarding ownership, copyright, and patentability of AI-generated works. As the whole field continues to advance, legal frameworks are also evolving to address the emerging challenges and opportunities.

C. Wisdom of the Crowd vs Epistemological Relativism

The responses generated by Generative AI, and LLMs in particular, are based on statistical patterns and associations learned from vast amounts of training data. While the models themselves do not have the ability to directly assess the majority opinion or conduct a voting process about what is true and what is false, the crowd-sourcing method used to collect the training data may rely on the previously described wisdom of the crowd principle. By its nature, it assumes that the sufficient number and diversity of contributors will ensure that individual biases and possible errors will be levelled out by the rest of the contributors. This can be seen as a strong assumption, especially from the epistemological relativism point of view. *Epistemological relativism* is a philosophical position, according to which knowledge and truth are not absolute, universal, or objective, but are instead relative to specific individuals, cultures, societies, or historical contexts [20]. In other words, different perspectives, beliefs, and interpretations can be equally valid and legitimate, depending on the context in which they arise. Epistemological relativism challenges the notion of universal truths and emphasises the role of individual perspectives and cultural contexts in shaping knowledge and truth. It focuses on the diversity of subjective interpretations. The wisdom of the crowd, on the other hand, suggests that aggregating diverse perspectives and independent judgements can lead to a single, more accurate outcome.

These two conflicting viewpoints again highlight the need for a thoroughly designed community-driven moderation framework in the proposed architecture. The outputs of LLMs are influenced by the distribution of the training data, inevitably containing certain biases and limitations. Even in the presence of a moderation framework, when using LLMs in applications involving decision-making or opinion representation, it is crucial to consider the limitations and potential biases in the training data and to supplement the outputs with appropriate human judgement and critical evaluation.

[5] https://www.melloddy.eu/

D. Bio-inspired Decision Making and Moderation

One possible way of enhancing the conventional AI moderation is to enhance it existing approaches from other relevant scenarios. Multiple crowd-sourcing contributors participating in a federated learning setup can be seen as individual agents providing their individual information into the global shared pool. Research approaches in the direction of multi-agent systems [21] developed many decentralised decision-making mechanisms inspired from malty-party auctions, arbitration, *etc.* Swarm intelligence [22] learns how advanced intelligence emerges from a swarm of low-intelligent individuals. Furthermore, as a step towards Artificial General Intelligence, community-driven Generative AI could also benefit from even more advanced mechanisms to address the challenge of decentralised decision making and moderation, and look into bio-inspired approaches and, more specifically, into how such processes are organised within a human brain.

Although it may sound counter-intuitive, a human brain is a highly decentralised system. A commonly accepted modern theory explains a brain's decision making based on a *global neuronal workspace* [23]. When individual inputs from the body arrive, multiple local processors within the brain autonomously perform continuous analysis of different parts of the input, also in combination with the local knowledge available to them. The results from these local processors are then 'broadcast' to a global neuronal workspace, from where other local processors can receive them. These other local processors may (or may not) find certain shared results interesting, pick them up for further evaluation and eventually place them back into the workspace. This way, certain results become more 'popular' than others, resulting in more and more local processors noticing and picking them up, and eventually becoming the final decision adopted by the whole brain. This process is to a great extent again similar to the majority and plurality principles observed in modern democracy. Applying this bio-inspired global workspace theory to the community-driven Generative AI for decentralised decision making and moderation might be an interesting and promising direction, which is however still at a very early stage of explorations [24]. Many fundamental challenges remain unaddressed, such as how to effectively broadcast local results, how to attract attention of relevant agents, how to evaluate and define the winning majority of the results, *etc.*

E. Federated Machine Unlearning

Within the sphere of community-driven AI, the elimination of unwelcome content already present in the trained models is of paramount importance. Sources of such undesirable content could be manifold. For example, adversarial data, when used in the training of generative AI models, can infuse inaccurate information into the resulting model. Likewise, adversarial attacks might leak confidential data kept within the model or users might occasionally produce data, such as inadvertent search queries leading to incorrect recommendations. Furthermore, community-driven AI systems, when trained on public datasets, often unintentionally inherit deep-seated racial and cultural biases. All such undesirable content contributes to encouraging biases, spreading harm, and eroding human dignity, and therefore should be wiped out from the models.

To this end, another important aspect of community-driven AI moderation is *unlearning* [25], [26] – *i.e.*, excluding some training results after they have already been included in the model. Federated machine unlearning represents the collective process of detecting sensitive and inaccurate predictions in a community-driven AI model and collaboratively unlearning the information housed within the model. This collaboration might commence with an individual's proposal to unlearn a category, a sample, a task, user-contributed data, or a data stream created by the AI. Once a proposal is tabled, the community should embark on an open and transparent process of consensus-building regarding what to unlearn and the specific methodology to follow. Every stage of the consensus-building journey for each proposal should be documented and retrievable. Upon reaching a consensus, new datasets for training should be constructed to revise what was previously learned. As an example, Wu *et al.* [27] delve into federated machine unlearning by reversing the stochastic gradient descent process for training and implementing elastic weight consolidation. After the fine-tuning/training phase, it is crucial to confirm and quantify the level of unlearning [28] achieved within the updated AI model. The data generated for 'forgetting' should challenge the model to reveal sensitive and incorrect information intended to be forgotten. Metrics should assess the degree of forgetting realised by the model in a recurring unlearning process. Federated machine unlearning is an innovative and emerging field where building consensus and unlearning present considerable challenges.

IV. Conclusion

In this paper, we aimed to challenge the establishing monopoly of the Big Tech on the Generative AI market by proposing a conceptual architecture for community-driven Generative AI. The envisioned architecture consists of several technological building blocks, among which we consider crowd-sourcing and federated learning to be the main enablers. By soliciting training data from a wide range of contributing parties, crowd-sourcing can capture a range of opinions, insights, and expertise that might otherwise be missed, thus providing a more comprehensive and unbiased view on a topic than any individual or organisation can offer. By drawing on a wider pool of perspectives and experiences this way, crowd-sourcing can help achieving the so-called wisdom of the crowd, where the collective intelligence arises from the aggregation of individual opinions, perspectives, and experiences, which can cancel out errors and biases and lead to more accurate and robust outcomes. At the same time, federated learning will ensure that data will remain private, since it assumes that instead of sending data samples to the central server, each client performs local training on its own data, and only exposes the updated model parameters which are then aggregated and shared back to the participating clients.

We have also considered several assumptions and potential research directions which still need further investigations. These include the important role of the community that will drive the whole development process, the IPR implications associated with the AI-generated content, the general epistemological concerns of the knowledge used to train the AI models, and finally possible ways of enhancing the AI moderation system by applying bio-inspired decision making and federated machine unlearning. Addressing all these open questions requires input from multiple stakeholders, including technologists, researchers, content creators, and ordinary users, as well as policy makers and civil society organisations.

ACKNOWLEDGMENT

This work has received funding from the European Union's HEU and H2020 research and innovation programmes under grant agreements No. 101095634 (ENTRUST) and No. 101020416 (ERATOSTHENES), and from the Research Council of Norway's BIA-IPN programme under grant agreement No. 309700 (FLEET).

REFERENCES

[1] V. Talla, M. Hessar, B. Kellogg, A. Najafi, J. R. Smith, and S. Gollakota, "LoRa Backscatter: Enabling The Vision of Ubiquitous Connectivity," *Proceedings of the ACM on interactive, mobile, wearable and ubiquitous technologies*, vol. 1, no. 3, pp. 1–24, 2017, doi: https://doi.org/10.1145/3130970.

[2] M. R. Ebling, "Pervasive Computing and the Internet of Things," *IEEE Pervasive Computing*, vol. 15, no. 1, pp. 2–4, 2016, doi: https://doi.org/10.1109/MPRV.2016.7.

[3] R. Dautov and S. Distefano, "Three-level hierarchical data fusion through the IoT, edge, and cloud computing," in *Proceedings of the 1st International Conference on Internet of Things and Machine Learning*. ACM New York, NY, USA, 2017, pp. 1–5, doi: https://doi.org/10.1145/3109761.3158388.

[4] R. Dautov, S. Distefano, D. Bruneo, F. Longo, G. Merlino, and A. Puliafito, "Pushing intelligence to the edge with a stream processing architecture," in *2017 IEEE International Conference on Internet of Things (iThings) and IEEE Green Computing and Communications (GreenCom) and IEEE Cyber, Physical and Social Computing (CPSCom) and IEEE Smart Data (SmartData)*. IEEE, 2017, pp. 792–799, doi: https://doi.org/10.1109/iThings-GreenCom-CPSCom-SmartData.2017.121.

[5] X. Wang, Y. Han, V. C. Leung, D. Niyato, X. Yan, and X. Chen, *Edge AI: Convergence of edge computing and artificial intelligence*. Springer, 2020, doi: https://doi.org/10.1007/978-981-15-6186-3.

[6] E. J. Husom, R. Dautov, A. Nedisan Videsjorden, F. Gonidis, S. Papatzelos, and N. Malamas, "Machine Learning for Fatigue Detection using Fitbit Fitness Trackers," in *Proceedings of the 10th International Conference on Sport Sciences Research and Technology Support - icSPORTS*, INSTICC. SciTePress, 2022, pp. 41–52, doi: https://doi.org/10.5220/0011527500003321.

[7] R. Dautov, E. J. Husom, F. Gonidis, S. Papatzelos, and N. Malamas, "Bridging the Gap Between Java and Python in Mobile Software Development to Enable MLOps," in *2022 18th International Conference on Wireless and Mobile Computing, Networking and Communications (WiMob)*. IEEE, 2022, pp. 363–368, doi: https://doi.org/10.1109/WiMob55322.2022.9941679.

[8] R. Dautov, S. Distefano, G. Merlino, D. Bruneo, F. Longo, and A. Puliafito, "Towards a Global Intelligent Surveillance System," in *Proceedings of the 11th International Conference on Distributed Smart Cameras*. ACM New York, NY, USA, 2017, pp. 119–124, doi: https://doi.org/10.1145/3131885.3131918.

[9] R. Dautov, S. Distefano, D. Bruneo, F. Longo, G. Merlino, A. Puliafito, and R. Buyya, "Metropolitan intelligent surveillance systems for urban areas by harnessing IoT and edge computing paradigms," *Software: Practice and experience*, vol. 48, no. 8, pp. 1475–1492, 2018, doi: https://doi.org/10.1002/spe.2586.

[10] J. Konečnỳ, H. B. McMahan, F. X. Yu, P. Richtárik, A. T. Suresh, and D. Bacon, "Federated learning: Strategies for improving communication efficiency," in *29th Conference on Neural Information Processing Systems (NIPS2016)*, 2016, pp. 1–5, doi: https://doi.org/10.48550/arXiv.1610.05492.

[11] I. Hegedűs, G. Danner, and M. Jelasity, "Gossip Learning as a Decentralized Alternative to Federated Learning," in *Distributed Applications and Interoperable Systems (DAIS 2019)*, J. Pereira and L. Ricci, Eds. Springer, 2019, pp. 74–90, doi: https://doi.org/10.1007/978-3-030-22496-7_5.

[12] G. Li, Y. Hu, M. Zhang, L. Li, T. Chang, and Q. Yin, "FedGosp: A Novel Framework of Gossip Federated Learning for Data Heterogeneity," in *2022 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*. IEEE, 2022, pp. 840–845, doi: https://doi.org/10.1109/SMC53654.2022.9945192.

[13] S. Boyd, A. Ghosh, B. Prabhakar, and D. Shah, "Gossip algorithms: Design, analysis and applications," in *Proceedings IEEE 24th Annual Joint Conference of the IEEE Computer and Communications Societies.*, vol. 3. IEEE, 2005, pp. 1653–1664, doi: https://doi.org/10.1109/INFCOM.2005.1498447.

[14] D. Shah, "Gossip algorithms," *Foundations and Trends® in Networking*, vol. 3, no. 1, pp. 1–125, 2009, doi: https://dx.doi.org/10.1561/1300000014.

[15] H. O. Ikediego, M. Ilkan, A. M. Abubakar, and F. V. Bekun, "Crowdsourcing (who, why and what)," *International Journal of Crowd Science*, vol. 2, no. 1, pp. 27–41, 2018, doi: https://doi.org/10.1108/IJCS-07-2017-0005.

[16] L. Duan, T. Kubo, K. Sugiyama, J. Huang, T. Hasegawa, and J. Walrand, "Incentive mechanisms for smartphone collaboration in data acquisition and distributed computing," in *2012 Proceedings IEEE INFOCOM*. IEEE, 2012, pp. 1701–1709, doi: https://doi.org/10.1109/INFCOM.2012.6195541.

[17] I. Kremer, Y. Mansour, and M. Perry, "Implementing the "wisdom of the crowd"," *Journal of Political Economy*, vol. 122, no. 5, pp. 988–1012, 2014, doi: https://doi.org/10.1086/676597.

[18] T. Gillespie, "Content moderation, ai, and the question of scale," *Big Data & Society*, vol. 7, no. 2, p. 2053951720943234, 2020, doi: https://doi.org/10.1177/2053951720943234.

[19] D. P. Anderson, J. Cobb, E. Korpela, M. Lebofsky, and D. Werthimer, "SETI@home: an experiment in public-resource computing," *Commun. ACM*, vol. 45, no. 11, pp. 56–61, 2002, doi: https://doi.org/10.1145/581571.581573.

[20] S. Luper, "Epistemic relativism," *Philosophical Issues*, vol. 14, pp. 271–295, 2004, doi: http://dx.doi.org/10.1111/j.1533-6077.2004.00031.x.

[21] A. Dorri, S. S. Kanhere, and R. Jurdak, "Multi-Agent Systems: A Survey," *IEEE Access*, vol. 6, pp. 28 573–28 593, 2018, doi: https://doi.org/10.1109/ACCESS.2018.2831228.

[22] A. Chakraborty and A. K. Kar, "Swarm Intelligence: A Review of Algorithms," in *Nature-inspired computing and optimization: Theory and applications*, S. Patnaik, X.-S. Yang, and K. Nakamatsu, Eds. Springer, 2017, pp. 475–494, doi: https://doi.org/10.1007/978-3-319-50920-4_19.

[23] G. A. Mashour, P. Roelfsema, J.-P. Changeux, and S. Dehaene, "Conscious Processing and the Global Neuronal Workspace Hypothesis," *Neuron*, vol. 105, no. 5, pp. 776–798, 2020, doi: https://doi.org/10.1016/j.neuron.2020.01.026.

[24] R. VanRullen and R. Kanai, "Deep learning and the global workspace theory," *Trends in Neurosciences*, vol. 44, no. 9, pp. 692–704, 2021, doi: https://doi.org/10.1016/j.tins.2021.04.005.

[25] H. Zhang, T. Nakamura, T. Isohara, and K. Sakurai, "A Review on Machine Unlearning," *SN Computer Science*, vol. 4, no. 4, p. 337, 2023, doi: https://doi.org/10.1007/s42979-023-01767-4.

[26] H. Xu, T. Zhu, L. Zhang, W. Zhou, and P. S. Yu, "Machine unlearning: A survey," *ACM Computing Surveys*, 2023, doi: https://doi.org/10.1145/3603620.

[27] L. Wu, S. Guo, J. Wang, Z. Hong, J. Zhang, and Y. Ding, "Federated Unlearning: Guarantee the Right of Clients to Forget," *IEEE Network*, vol. 36, no. 5, pp. 129–135, 2022, doi: https://doi.org/10.1109/MNET.001.2200198.

[28] X. Gao, X. Ma, J. Wang, Y. Sun, B. Li, S. Ji, P. Cheng, and J. Chen, "VeriFi: Towards Verifiable Federated Unlearning," *Computing Research Repository*, 2022, doi: https://doi.org/10.48550/arXiv.2205.12709.

Position Papers of the 18th Conference on Computer Science and Intelligence Systems pp. 51–58
DOI: 10.15439/2023F5947
ISSN 2300-5963 ACSIS, Vol. 36

Federated Learning for Data Trust in Logistics

Michael Koch
Institute for applied Computer Science (InfAI) e.V.
Goerdelerring 9, 04109 Leipzig, Germany
Email: koch@infai.org

Sascha Kober, Stanislaw Straburzynski, Benjamin Gaunitz, Bogdan Franczyk
Leipzig University
Faculty of Economics
Grimmaische Straße 12
04109 Leipzig, Germany
Email: {kober, straburzynski, gaunitz, franczyk}@wifa.uni-leipzig.de

***Abstract*—In the field of logistics, there is a significant shortage of qualified employees. Artificial Intelligence (AI) can help solve that problem supporting existing employees and reducing their workload. However, large amounts of data to train AI models are required and, in most cases, due to lack of trust between companies, model training is based solely on locally stored data from logistics providers and some publicly available datasets. To address this data scarcity issue, a proposed solution is to employ federated learning (FL), in the context of data trust (DT) by training AI models across multiple companies, based on both centralized data, within the DT platform and decentralized data from logistics providers data silos, while ensuring data sharing access at the attribute level. This paper proposes this approach and points out the importance of data sharing for effective model training for solving workforce challenges in logistics.**

***Index Terms*— data trust, federated learning, logistics, machine learning, artificial intelligence**

I. Introduction

According to Germany's domestic freight transport statistics in 2021, 37.6% of transport vehicles were empty at every driven kilometer, a recurring trend as in previous years [1]. Additionally, the degree of utilization of loading capacity and transport performance has decreased from 40% in 2002 to 34.7% in 2021 and 45.9% to 41.4% respectively [1]. The degree of utilization of loading capacity is defined as how full a vehicle is in relation to its total loading capacity. The transport performance is a statistic that summarizes various key figures for freight transport. At the European Union level, one in every five kilometers is travelled by an empty vehicle in 2020 and an average of 24% of national transport is empty [2], [3]. Another problem is the shortage of drivers and logistics workers with high percentages of unfilled positions in various roles [4]. The logistics industry also faces issues related to communication, collaboration, flexibility in capacity planning [5] and the lack of reliance in existing online platforms [6].

To address these challenges, the use of data trust (DT) platforms as secure and transparent platforms for data exchange and storage [6] has been proposed. This DT is a neutral entity that serves a DT ecosystem for data exchange and is managed by a transparent and non-profit organization. This DT should implement state-of-the-art data security techniques that meet the requirements of logistics companies, including efficient access and usage control concepts at the attribute level, based on the logistics business data. This paper builds upon the ongoing TRANSIT project at the University of Leipzig in exchange with the participating logistics providers [7], as well as on research in current freight exchange platforms [8].

The idea is to support logistics service providers with artificial intelligence (AI) in the context of a DT platform. This can be done by providing incentives or creating a benefit for platform usage and thus addressing the shortage of qualified workers [9]. Examples of AI support include assisting workers in calculating transport prices, which is currently a complex process or helping them to find potential cooperation partners, because many of these processes are manual and based on the knowledge of just a few employees within a company. To achieve accurate results, a substantial amount of business data from each logistics provider is necessary [10]. However, this data are primarily stored in the local data silos of the logistics company rather than the DT platform. However, many companies prefer not to upload data to DT platform. Data minimization in

The position paper is funded by Federal Ministry of Education and Research (Project ID: 16DTM109B) and by the European Union - NextGenerationEU.

©2023, PTI

Thematic track: Distributed Edge AI – Risks and Challenges

the General Data Protection Regulation (GPDR) [11] and potentially much larger historical data can be among the reasons.

Leveraging this vast source of data, without transporting and transforming it within the platform, is a challenge, but it increases the trust of logistics companies. The utilization of federated learning (FL) and efficient usage control for the external data stores provides a solution within a DT [12]. This solution trains the models in each of the data silos. Therefore, a comparable input format is required. Then the data trustee will act as a neutral authority at this point, advising the logistician on how to transform the data locally to obtain high-quality training data for any training purpose. The combination of DT and FL has the following advantages:

- Reducing the effort to merging, combining, and normalizing the local heterogeneous data
- High scalability on distributed and heterogeneous hardware
- Enhanced security, as there is no direct access to the locally stored data
- Enrichment of the data with publicly available data

To leverage this large amount of data and achieve a critical mass of users, it is crucial to train AI models securely and fairly [13]. A significant research topic in FL is ensuring fairness in the returned results obtained by users based on the quantity and quality of data used for training the model, as well as its impact on global model parameters and gradients. This idea is particularly relevant to the logistics sector, where established companies do not want to share the information which they have collected over the years and which shows some internals, like the price calculation.

Therefore, an effective mechanism for training and maintaining the model is important when considering scenarios where start-ups want to participate with their limited data or attackers want to obtain a data leak that reveals business secrets. For this proposal, the central global model should not predict exact values. Also, the parameters of the shared model should be noisy so that it is impossible to obtain data used during model training. Furthermore, the predictions and the downloaded model should be adjusted in accuracy based on the amount of data that a company has provided to the model training.

By adopting this approach, the platform can increase its usage, efficiently and effectively address the shortage of skilled workers, and facilitate better management and collaboration. The utilization of federated AI to consider various aspects of the market through a combined overall model, such as preventing price dumping by calculating fair market prices, provides benefits to all companies and will increase the usage of the platform.

II. Related Work and Background

This section provides an overview of related work and research that impacts on successful realization of the proposed solution. Topics which will be related to problem-solving are DT platforms, access, and usage control and at least FL.

A. Data trust

The concept of Data Trust (DT) has been introduced as a means of facilitating the exchange of data between different entities [6]. This DT can be imagined as a neutral entity, such as a not-for-profit organization or a transparent company, that establishes a data sharing ecosystem. This instance provides the infrastructure for data sharing, which can be designed in different ways, such as DT as a service [13].

The actors defined in the context of DT are the data trustee, data provider and data user. The data trustee is responsible for providing the above-mentioned infrastructure. The data provider, which can be a company or person, contributes the data for sharing. This can involve transmitting the data to the platform and storing it there or transmitting it directly to the data user while the metadata from the data source are stored securely on the DT.

This data user, represents persons and companies that utilize the provided data to develop innovative products for the data provider or other clients. However, they are not authorized to sell the data without the permission of the data provider, and they can only process the data on the terms and conditions agreed with the data provider.

Managing a DT involves considering a number of aspects, including internal governance, user interaction and market structure, which are introduced in [14]. Legal regulations, such as GDPR and Data Governance Act (DGA), also are related to this issue when managing data within a platform or as part of a DT [15], [11]. One approach, introduced by Lomotey et al. [13] is DT as a Service, wherein the data trustee deploys a DT platform for a specific application domain and integrates services into the platform on demand.

It is important to note that the data trustee does not generate revenue by handling the data. Instead, the DT finances itself through user license fees or providing additional services in consultation with the data providers.

B. Access and Usage Control

Access and usage control are essential for ensuring data security within the platform. This fine-grained access control ensures that data providers can define access permissions at the attribute level of entities, such as the street of an address. It should then support centralized and decentralized scenarios, to enable secure data sharing between logistic service providers and researchers. To fulfill these requirements, access and usage control concepts need to be

TABLE I.
ACCESS CONTROL

concept	short description	Source
IBAC	identity-based	[16]
RBAC	role-based	[17]
ABAC	attribute-based	[16], [18]
ReBAC	relation-based	[19]
UCON	usage control	[20]

analyzed, evaluated, and continuously developed at the conceptual and implementation levels. One recommended access control architecture for the underlying DT platform is the zero-trust model [16] where each data access request is thoroughly evaluated and can be combined with different access control mechanisms.

Table 1 provides an overview of existing access control concepts. Through preliminary research, literature review and analysis of existing software, it became apparent that a specific problem could not be solved directly and required workarounds. A major logistics requirement, which relates to setting permissions at the attribute level, is often addressed by mapping at the entity level, where a unique identifier exists. In addition, certain parts of the data may need to be shared with all platform participants, such as when seeking collaboration partners, while ensuring that certain government agencies have full access to a company's data. It is additionally important for the data provider to be able to distinguish and configure which individuals, or companies can access their data.

Furthermore, in the context of implementation of artificial intelligence in the DT platform, there is a need for use control to regulate how data can be processed. In this way, it is possible for all parties involved in logistics to determine which models can use their data for training purposes. The preferred access model is Attribute-Based Access Control (ABAC) because it can also incorporate Identity-Based Access Control (IBAC) and Role-Based Access Control (RBAC) as attributes of a sign-on user, providing a higher level of granularity. Relation-based access control (ReBAC) and usage control (UCON) are suitable for more complex use cases and are not currently required for this DT implementation. The advantage of using ReBAC is to define access on an flexible data model [17] and allow clean and fast retrieval of access rights, but in actual implementation it needs a unique identifier for all entities, which is not always present in logistics data exchange, as they also share data at the attribute level.

C. *Federated Learning*

FL is an approach to AI in which the data are kept in local data silos, and the algorithm is applied to those silos for training purposes. However, the data must be transformed locally into a comparable input format before the training. The trained models can then be merged either centrally or in a decentralized manner. This concept was first introduced by Mahan et al. [18] and involves multiple iterative steps in the training process including client selection, client computation, model aggregation, model update and convergence checking.

In the client selection step, the coordinator or an algorithm in the P2P network chose clients based on various criteria. These criteria include historical activity, such as previous involvement in the training process or computation time, as well as factors such as model quality, influence on the global model, quality of training data, and technical characteristics such as network bandwidth and memory.

The calculation step is then executed on the selected clients. They receive the global model trained in previous training rounds or initialized with gradients and execute training with their local data and the specified computational parameters.

The next step involves aggregating of the locally trained models. Multiple approaches for model aggregation exist, which may include verification of the model on cybersecurity, etc. [23]. To address privacy concerns, concepts such as Differential Privacy (DP) or Homomorphic Encryption (HE) can be applied before transmitting the local model [24]. On the one hand, this approach provides the advantage of securing locally sensitive data. On the other hand, there are challenges such as longer convergence times in certain cases of high privacy with DP and increased computational overhead using HE. In addition, there is a commonly used approach where not all selected clients need to send their updated model; only the majority of clients participate in the aggregation.

In the fourth step, the coordinator or selected clients in a decentralized environment updates the model. A variety of algorithms can be used for this update. These include gradient averaging and optimization techniques such as the Adam optimizer [24]. The aim is to speed up the convergence process and to improve the accuracy of the resulting model. It is also beneficial to adopt a secure aggregation protocol to enhance cyber security [24], but at the cost of increased communication overhead.

The last step is to verify that the convergence criteria have been met, typically by evaluating whether the convergence error is below a predefined threshold. The purpose of this step is to ensure that the training process has achieved a satisfactory level of accuracy of the resulting model. If the algorithm does not converge, it restarts from the selection step.

In the field of FL, recent studies have identified four main research directions [10], [25], [26]: cybersecurity, fairness, optimization of aggregation and computation of heterogeneous FL [10] and the analysis of these points in the context of blockchain technology [27].

Attempts have been made to combine these research directions and design a comprehensive framework or FL algorithm that addresses openness, security, fairness, and decentralization. Such a successful attempt can be found in

[28], which entails further testing and analysis in logistics assuming a DT environment.

Recent directions in FL explore centralized pre-training of models to improve convergence and accuracy [29]. In this setting, proxy data do not present the same privacy concerns as in a DT scenario, so that privacy mechanisms such as DP must also be applied at this point. Adaptive central training methods in FL for faster convergence have also been proposed [30]. Furthermore, there is ongoing research on the combination of FL and central learning, which is called mixed FL (MFL) [31], [32]. Until now, MFL has focused on analyzing independent and identically distributed (iid) settings, especially in the horizontal FL (HFL) approach, to achieve better results in terms of convergence, accuracy, communication time, and cost [31]. However, for both federated and centralized data in the context of a DT, these algorithms have yet to be fully considered from a privacy and fairness perspective.

Furthermore, hybrid FL (HFL) has been introduced as another approach that combines HFL and vertical FL (VFL) [33] and shows promise for future work. HFL and VFL each refer to data distribution and require different algorithms. The presence of both data distributions and other federated learning paradigms requires evaluation to determine their potential benefits in terms of convergence time and accuracy. Hetero FL [34] is another approach that allows training a global model based on different local model architectures. This is particularly relevant for logistics DT, where data access is handled at the attribute level, which may be different for each logistics service provider.

Regarding privacy, there are ideas for knowledge transfer [35], with research focusing primarily on public centralized data and transferring knowledge from the locally trained models using one-way knowledge transfer.

Despite these research directions, there are still open issues that need to be addressed. Kairouz et al. [36] examine various challenges associated with federated learning (FL), including the context of cross-silo FL and fairness. The term "cross-silo" refers to the process of training models across multiple organizations, each managing large data silos. These issues include the absence of certain features due to varying data sharing rules and incomplete data entry, which can affect data trust. Other problems are difficulties in data normalization, like different storage formats and inconsistency of data and labels, differences in privacy policies among logistics service providers, and fairness concerns regarding the selection of training participants based on available hardware and further features.

Many of them can be still optimized, validated and further developed in the context of DT environment. In addition, there is a need to explore the combination of the HFL and MFL approaches, which includes the basic distribution of the data to a data trustee in the logistics domain.

III. PROTOTYPE

For the prototype, a new access and usage control mechanism will be implemented first to fulfill the requirements of a secure DT with specific requirements in logistics. Once this implementation is complete, the following stage involves requirements engineering, implementation and deployment of the FL framework on a scalable engine, such as Docker [37].

A. Access Control for data trustees in logistics

To meet the requirements of protecting and sharing data at the attribute level, for both internal and external data, the access control model should be designed using state-of-the-art techniques.

An access control mechanism based on ABAC, in more specific Next Generation Access Control (NGAC), should be implemented to secure the internal data and to be able to share it with other logistics service providers and researchers. Through an access control system for FL, researchers only have access to the in- and output of the FL Framework and not directly to the data.

This access control will be implemented as an extension of the Policy Machine [38], in which NGAC is already integrated. This framework can also be used with other access mechanisms such as RBAC. However, fine-grained data sharing at the attribute level, which is essential for logistics, has not yet been integrated.

Figure 1 illustrates the initial design draft of the access data model, which will serve as a repository for the storage of the access rules. In addition to this model, there are four group categories and several further restrictions. These groups are categorized as *standard, FL-internal, FL-external* and *any*. Considering that the platform is designed for logistics, the term "company" is used in the following text to refer to these groups. The *standard* company represents the logistic companies themselves. The *FL-internal* company consists of the platform provider and those who require access to the data for the FL training. They can get it through the FL control access model via a search API, which is only available to specific individuals or for internal data exchange to the FL framework. The *FL-external* company is formed by researchers who are allowed to train AI models by executing jobs through an API, but they are not permitted to have direct access to the data. There is always a mandatory privacy measure for trained models, with minimal privacy requirements for AI models returned to this company. The fourth company is called *any* and consists of all company members who also have access to the *standard* company type.

The user’s access is determined by the group to which they belong, with additional consideration given to roles within *standard* and *any* company. If the data come from the user's company or has been shared, there is a second

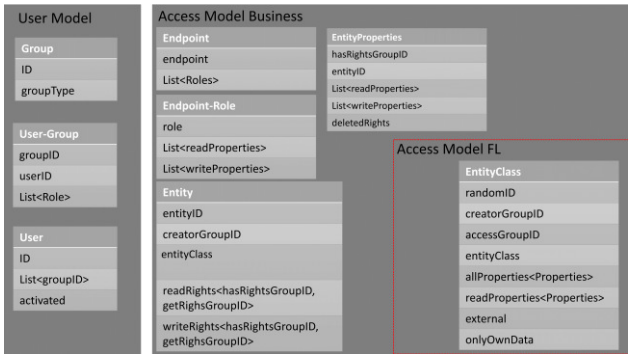


Fig. 1 Access Data Model – initial draft

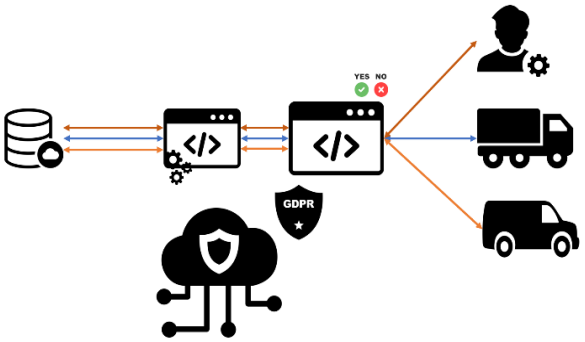


Fig.2 Concept of DT. Icons are from [42].

level of access control. In terms of training, FL algorithms have two levels of access. They can either use only a company's own data, or they can access all shared data from companies. This works according to the permissions granted by the logistics service providers. This two-level access provides better control over data usage and provides an incentive for the platform. The next step involves the transformation in the NGAC access schema and extends it with the attribute level and metadata access control schema.

B. Data trust in logistics

The DT is a web service designed to secure and manage logistics company data. Direct access to the data is restricted by the participating companies and can be shared with other companies or utilized as training input in the FL. An overview of data trust is shown in Fig. 2 and the FL concept is shown in Fig. 3. Fig. 2 shows the companies, the data providers and users and the data trustee, who have access to the data and the processing unit based on the access control rules.

To increase participants' trust in the platform, the DT incorporates state-of-the-art techniques such as Zero Trust Architecture [16] with micro-segmentation, which brings the access control mechanism closer to the data source [17]. The DT also aims to encourage companies to provide their own data for research purposes. The first incentive is that the platform provider acts as a data trustee under a non-profit organization. The second stimulus is the potential benefit of utilizing a federated pricing model or being recommended as a potential collaboration partner if a company shares its data for FL. Another important aspect to consider is fairness in terms of prediction accuracy based on the data provided. The fourth encouragement concerns the data available to researchers, who will only have access to the FL framework and not directly to the data.

For protection against privacy attacks on AI models, any external access to the model should be subject to a mandatory privacy mechanism. These trained models can then be utilized by data scientists and companies for a variety of use cases, such as numerous pricing models or other predictions, such as time optimization or CO2 reduction. There are also ongoing research efforts exploring new data platforms [39], model input parameter selection [40] and pricing strategies in competition between new and existing logistics service providers [41].

Furthermore, depending on the role within the AI ecosystem and whether the data is provided by companies for training or only used by researchers, there should be control over the use of the trained models, as well as the accuracy and privacy of the predictions.

C. FL-DT

Based on the access control model of the metadata for AI training, an FL training framework will be developed. The architectural concept is illustrated in Fig. 3, where companies, the data trustee and researchers have access to the FL training environment, secured by access control. In this example, a company participates in the training with its local data. The infrastructure for the platform and the FL framework is hosted on a scalable framework network.

The first module is data preprocessing. This is where the data are transformed or pre-processed to get the local data in a comparable input format. For this purpose, the data are transferred to a " trusted environment". This environment can be hosted either at the logistics provider, in the platform on the scalable environment, or at a provider trusted by the logistics provider. The logistician decides where this environment exists.

It is like a container that takes care of data preparation and model training. Within the „trusted environment," it is now possible to transform the data, which can be done by the logistician itself or in exchange with the neutral entity of the data trustee, to obtain high quality training data. Additionally, the data can be scaled using a pre-configured and secure scaler. This scaler can be pre-trained solely on the central data from the same company or with privacy additions from all companies. At this point, it should also be possible to tag the imported data with release numbers for reproducibility and explainability. It should also be possible to delete a release if the logistician wishes to do so or regulatory rights require this.

Once this data has been provided, the training phase can commence. This involves addressing various gaps that are specific to DT in logistics and FL. Due to the secure infrastructure approach, all data access and exchange pass

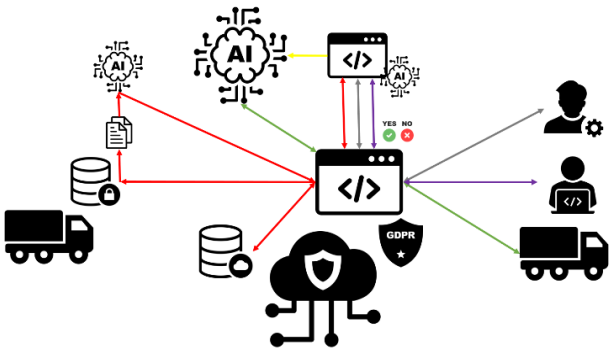


Fig. 3 Concept of FL in data trust. Icons are from [42].

through the access control module, ensuring authentication and authorization. To address this privacy concerns during the training phase, privacy mechanisms such as DP or one-time passwords be employed for the model's data exchange. At this point, blockchain technology would be avoided. This is due to transactional time limitations and legal requirements, such as the right to erasure outlined in the General Data Protection Regulation (GDPR), where only some research idea results are available [43], which are not yet as effective as they need to be.

The model initialization and pretraining is the first step and different approaches exist based on security and fairness considerations. One solution is to pre-train a central model for faster convergence [29], but in the context of DT with sensitive data and no public data. The second approach is to use MFL. The question of which FL approach can help to overcome the challenges posed by different data distributions and sharing combinations in the logistics domain is crucial at this step.

The second step involves training of the model. In this step, the model is transmitted to the "trusted environment" where the pre-processed local data are stored. During the training phase, only the model itself and the training parameters are exchanged with this environment. It is important to implement security measures to ensure that the model works well, while considering the security implications for the participants. One approach to achieve this is the utilization of Differential Privacy (DP).

During training it may be possible to use alternative models, due to different attribute sharing levels between the companies. This could work similarly to transfer learning, whereby the model first trained to the same output error as the central model and then further trained on local data. If this local model works better for their specific use cases, it can be saved and accessed exclusively by the company.

After the training phase, the models are returned to the center where the models are combined by various algorithms or where FL-PATE [35], a knowledge transfer algorithm, would be used. Subsequently, another set of central data is then utilized to train the model.

The next step involves testing whether the model convergence criteria have met. If the criteria are satisfied, the central model is stored in a data repository accessible to logistics service providers for their predictions, such as the delivery price. The accuracy will be different, or a range of output values will be predicted based on the quality and quantity of data supplied by the company for FL training. Researchers can download their trained models based on their chosen configurations or algorithms, with an additional privacy budget, such as DP, to ensure privacy preservation.

IV. Future Work

This paper presents a concept of the FL paradigm that is based on a DT in the logistics domain. Implementing the concept of FL framework in a scalable, secure, fair, and adaptable manner is aimed at in future work. Therefore, the focus would be on the following points that should be considered and researched.

A pre-research topic is to discover pricing models, as discussed in [44], and to find methods to automatically select relevant input attributes based on the secure DT and FL approach.

Furthermore, MFL and pre-trained FL be explored for faster convergence and higher accuracy with secure training. Next, HFL is an important topic for this FL approach. This happens because there are horizontal data across companies. On the other hand, there is also vertical data, such as shared orders, which are provided in diverse ways by the two companies involved in an FL training session. The approach must also be developed so that all shared data can be effectively incorporated into the training of the AI in accordance with the data sharing rules. For the central data in MFL, where the central data is also secret, a secure paradigm for the training process must also be used to ensure secure computation. For this task, it is necessary to be GPDR compliant, e.g., with secure aggregation [24]. Then fairness principles are required, to avoid over-representation of a single company in the AI prediction. Furthermore, exploring alternative incentives for companies to share their data is also essential and how these can be incorporated into the model training.

Considering asynchronous, decentralized, or hierarchical FL approaches in the context of a DT can be beneficial to achieve faster convergence and avoid single points of failure, which is also a part of further research. Finally, a scalable approach should be developed that works in a cluster setup for scalability.

This implementation will be empirically evaluated based on real data from small companies in Saxony, Germany, which, as usual, are interested in receiving an incentive and providing their business data for this purpose.

References

[1] BMDV. “Amtliche Güterkraftverkehrsstatistik.” https://www.kba.de/DE/Statistik/Kraftverkehr/deutscherLastkraftfahrzeuge/vd_Inlandsverkehr/vd_inlandsverkehr_node.html (accessed Apr. 18, 2023).

[2] Eurostat. “Summary of annual road freight transport by type of operation and type of transport (1 000 t, Mio Tkm, Mio Veh-km).” https://

www.eea.europa.eu/ds_resolveuid/2952faa0aff24c37aa0cfab6a86730c8 (accessed Apr. 20, 2023).
[3] "A fifth of road freight kilometres by empty vehicles," Eurostat, 12 Oct., 2021. https://ec.europa.eu/eurostat/web/products-eurostat-news/-/ddn-20211210-1 (accessed: Apr. 20, 2023).
[4] A. Streim and B. Kokott. "In der Logistik werden die Sicherheitsmaßnahmen verschärft." https://www.bitkom.org/Presse/Presseinformation/Digitalisierung-Logistik (accessed Apr. 14, 2023).
[5] C. Kille, T. Schmidt, W. Stölzle, L. Häberle, and S. Rank. "Begegnung von Kapazitätsengpässen im Straßengüterverkehr – Fokus Personal." http://logistik-digitalisierung.de/ (accessed Apr. 18, 2023).
[6] W. Hall and J. Pesenti, "Growing the artificial intelligence industry in the UK," 2017. [Online]. Available: https://www.gov.uk/government/publications/growing-the-artificial-intelligence-industry-in-the-uk
[7] [7] Insititut für Wirtschaftsinformatik – Universität Leipzig, TRANSIT Data trust for logistics: Data Trusts for Enhancing Logistics Collaboration. Accessed: May 22, 2023. [Online]. Available: https://transit-project.de/
[8] J. Witkowski, "Electronic Freight exchange and logsitics platforms in buildung of supply chains," 2018. [Online]. Available: https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=&ved=2ahUKEwi734WPh8L-AhXQp6QKHSo9BdwQFnoECAgQAQ&url=https%3A%2F%2Fwww.confer.cz%2Fclc%2F2018%2Fdownload%2F2448-european-electronic-freight-exchange-as-a-future-central-coordinator-in-supply-chains.pdf&usg=AOvVaw38zq7w-pkr1klk3gvC12Vu
[9] K. Houser and J. W. Bagby, The Data Trust Solution to Data Sharing Problems, 2022.
[10] B. Liu, N. Lv, Y. Guo, and Y. Li, "Recent Advances on Federated Learning: A Systematic Survey," Jan. 2023, http://dx.doi.org/10.48550/arXiv.2301.01299. [Online]. Available: http://arxiv.org/pdf/2301.01299v1
[11] S. Stalla-Bourdillon, G. Thuermer, J. Walker, L. Carmichael, and E. Simperl, "Data protection by design: Building the foundations of trustworthy data sharing," Data & Policy, vol. 2, 2020, http://dx.doi.org/10.1017/dap.2020.1.
[12] X. Zhang, "A commentary of Data trusts in MIT Technology Review 2021," Fundamental Research, vol. 1, no. 6, pp. 834–835, 2021, http://dx.doi.org/10.1016/j.fmre.2021.11.016.
[13] R. K. Lomotey, S. Kumi, and R. Deters, "Data Trusts as a Service: Providing a platform for multi-party data sharing," International Journal of Information Management Data Insights, vol. 2, no. 1, p. 100075, 2022, http://dx.doi.org/10.1016/j.jjimei.2022.100075.
[14] A. Blankertz. "Designing Data Trusts: Why We Need to Test Consumer Data Trusts Now." https://www.stiftung-nv.de/sites/default/files/designing_data_trusts_e.pdf (accessed Apr. 20, 2023).
[15] C. L. Geminn, P. C. Johannes, J. K. M. Müller, and M. Nebel, Data Governance in Germany – An Introduction. Universität Kassel, 2023. [Online]. Available: https://kobra.uni-kassel.de/handle/123456789/14590
[16] S. Rose, O. Borchert, S. Mitchell, and S. Connelly, "Zero Trust Architecture," National Institute of Standards and Technology, 2020. http://dx.doi.org/10.6028/NIST.SP.800-207. [Online]. Available: https://csrc.nist.gov/publications/detail/sp/800-207/final?ref=hackernoon.com
[17] Ruoming Pang et al., "Zanzibar: {Google's} Consistent, Global Authorization System," in Proceedings of the 2019 USENIX Annual Technical Conference: July 10-12, 2019, Renton, WA, USA, 2019, pp. 33–46. [Online]. Available: https://www.usenix.org/conference/atc19/presentation/pang
[18] H. B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y. Arcas, "Communication-Efficient Learning of Deep Networks from Decentralized Data," http://dx.doi.org/10.48550/arXiv.1602.05629. [Online]. Available: http://arxiv.org/pdf/1602.05629v4
[19] N. F. Syed, S. W. Shah, A. Shaghaghi, A. Anwar, Z. Baig, and R. Doss, "Zero Trust Architecture (ZTA): A Comprehensive Survey," IEEE Access, vol. 10, pp. 57143–57179, 2022, http://dx.doi.org/10.1109/access.2022.3174679.
[20] R. S. Sandhu, E. J. Coyne, H. L. Feinstein, and C. E. Youman, "Role-based access control models," Computer, vol. 29, no. 2, pp. 38–47, 1996, http://dx.doi.org/10.1109/2.485845.
[21] D. F. Ferraiolo, R. Chandramouli, V. C. Hu, and D. R. R. Kuhn, "A Comparison of Attribute Based Access Control (ABAC) Standards for Data Service Applications," 2016, http://dx.doi.org/10.6028/NIST.SP.800-178.
[22] Ruoming Pang et al., "Zanzibar: Google's Consistent, Global Authorization System," 2019. [Online]. Available: https://www.semanticscholar.org/paper/Zanzibar%3A-Google's-Consistent%2C-Global-Authorization-Pang-C%C3%A1ceres/1362dec32d9d0b9d8b369f7ebcfef19bbc975066
[23] R. Sandhu and J. Park, "Usage Control: A Vision for Next Generation Access Control," in vol. 2776, 2003, pp. 17–31, http://dx.doi.org/10.1007/978-3-540-45215-7_2.
[24] N. Truong, K. Sun, S. Wang, F. Guitton, and Y. Guo, "Privacy Preservation in Federated Learning: An insightful survey from the GDPR Perspective," Nov. 2020, http://dx.doi.org/10.48550/arXiv.2011.05411. [Online]. Available: https://arxiv.org/pdf/2011.05411
[25] J. Zhang, H. Zhu, F. Wang, J. Zhao, Q. Xu, and H. Li, "Security and Privacy Threats to Federated Learning: Issues, Methods, and Challenges," Security and Communication Networks, vol. 2022, pp. 1–24, 2022, http://dx.doi.org/10.1155/2022/2886795.
[26] M. Alazab, S. P. RM, P. M, P. K. R. Maddikunta, T. R. Gadekallu, and Q.-V. Pham, "Federated Learning for Cybersecurity: Concepts, Challenges, and Future Directions," IEEE Trans. Ind. Inf., vol. 18, no. 5, pp. 3501–3509, 2022, http://dx.doi.org/10.1109/TII.2021.3119038.
[27] J. Zhu, J. Cao, D. Saxena, S. Jiang, and H. Ferradi, "Blockchain-empowered Federated Learning: Challenges, Solutions, and Future Directions," ACM Comput. Surv., vol. 55, no. 11, pp. 1–31, 2023, http://dx.doi.org/10.1145/3570953.
[28] G. Yu et al., "IronForge: An Open, Secure, Fair, Decentralized Federated Learning," Jan. 2023, http://dx.doi.org/10.48550/arXiv.2301.04006. [Online]. Available: https://arxiv.org/pdf/2301.04006
[29] J. Nguyen, J. Wang, K. Malik, M. Sanjabi, and M. Rabbat, "Where to Begin? On the Impact of Pre-Training and Initialization in Federated Learning," Oct. 2022, http://dx.doi.org/10.48550/arXiv.2210.08090. [Online]. Available: https://arxiv.org/pdf/2210.08090
[30] S. Reddi et al., "Adaptive Federated Optimization," Feb. 2020, http://dx.doi.org/10.48550/arXiv.2003.00295. [Online]. Available: https://arxiv.org/pdf/2003.00295
[31] S. Augenstein et al., "Mixed Federated Learning: Joint Decentralized and Centralized Learning," May. 2022, http://dx.doi.org/10.48550/arXiv.2205.13655. [Online]. Available: https://arxiv.org/pdf/2205.13655
[32] K. Yang, S. Chen, and C. Shen, "On the Convergence of Hybrid Server-Clients Collaborative Training," IEEE J. Select. Areas Commun., vol. 41, no. 3, pp. 802–819, 2023, http://dx.doi.org/10.1109/JSAC.2022.3229443.
[33] X. Zhang, W. Yin, M. Hong, and T. Chen, "Hybrid Federated Learning: Algorithms and Implementation," Dec. 2020, http://dx.doi.org/10.48550/arXiv.2012.12420. [Online]. Available: https://arxiv.org/pdf/2012.12420
[34] E. Diao, J. Ding, and V. Tarokh, "HeteroFL: Computation and Communication Efficient Federated Learning for Heterogeneous Clients," Oct. 2020, http://dx.doi.org/10.48550/arXiv.2010.01264. [Online]. Available: https://arxiv.org/pdf/2010.01264
[35] Y. Pan, J. Ni, and Z. Su, "FL-PATE: Differentially Private Federated Learning with Knowledge Transfer," in 2021 IEEE Global Communications Conference (GLOBECOM), 2021, http://dx.doi.org/10.1109/globecom46510.2021.9685079.
[36] P. Kairouz et al., "Advances and Open Problems in Federated Learning," Dec. 2019, http://dx.doi.org/10.48550/arXiv.1912.04977. [Online]. Available: http://arxiv.org/pdf/1912.04977v3
[37] Dirk Merkel, Docker: lightweight linux containers for consistent development and deployment. Houston, TX: Belltown Media, 2014. [Online]. Available: https://dl.acm.org/doi/10.5555/2600239.2600241
[38] D. Ferraiolo, V. Atluri, and S. Gavrila, "The Policy Machine: A novel architecture and framework for access control policy specification and enforcement," Journal of Systems Architecture, vol. 57, no. 4, pp. 412–424, 2011, http://dx.doi.org/10.1016/j.sysarc.2010.04.005.
[39] Y.-A. Du, "Research on the Route Pricing Optimization Model of the Car-Free Carrier Platform Based on the BP Neural Network Algorithm," Complexity, vol. 2021, pp. 1–10, 2021, http://dx.doi.org/10.1155/2021/8204214.
[40] M. Poliak, A. Poliakova, L. Svabova, N. A. Zhuravleva, and E. Nica, "Competitiveness of Price in International Road Freight Transport,"

JOC, vol. 13, no. 2, pp. 83–98, 2021, http://dx.doi.org/10.7441/joc.2021.02.05.

[41] F. Du, S. Ang, F. Yang, and C. Yang, "Price and distribution range of logistics service providers considering market competition," APJML, vol. 30, no. 4, pp. 762–778, 2018, http://dx.doi.org/10.1108/APJML-09-2017-0208.

[42] UXWing.com, Exclusive collection of free icons download for commercial projects without attribution. [Online]. Available: https://uxwing.com

[43] E. Politou, F. Casino, E. Alepis, and C. Patsakis, "Blockchain Mutability: Challenges and Proposed Solutions," Jul. 2019, http://dx.doi.org/10.48550/arXiv.1907.07099. [Online]. Available: https://arxiv.org/pdf/1907.07099

[44] H.-S. Jang, T.-W. Chang, and S.-H. Kim, "Prediction of Shipping Cost on Freight Brokerage Platform Using Machine Learning," Sustainability, vol. 15, no. 2, p. 1122, 2023. http://dx.doi.org/10.3390/su15021122. [Online]. Available: https://www.mdpi.com/2071-1050/15/2/1122

Position Papers of the 18th Conference on Computer Science and Intelligence Systems pp. 59–66
DOI: 10.15439/2023F8069
ISSN 2300-5963 ACSIS, Vol. 36

Comparison of Deep Learning Architectures for three different Multispectral Imaging Flow Cytometry Datasets

Philippe Krajsic
0000-0002-1722-8752
Leipzig University
Information Systems Institute
Grimmaische Straße 12, 04109 Leipzig, Germany
krajsic@wifa.uni-leipzig.de

Thomas Hornick, Susanne Dunker
0000-0003-0280-9260, 0000-0001-7276-776X
German Centre for Integrative Biodiversity Research
Department of Physiological Diversity
Puschstraße 4, 04103 Leipzig, Germany
thomas.hornick@idiv.de, susanne.dunker@idiv.de

***Abstract*—Multispectral imaging flow cytometry (MIFC) is capable of capturing thousands of microscopic multispectral cell images per second. Deep Learning Algorithms in combination with MIFC are currently applied in different areas such as classifying blood cell morphologies, phytoplankton cells of water samples or pollen from air samples or pollinators. The goal of this work is to train classifiers for automatic and fast processing of new samples to avoid labor-intensive and error-prone manual gating and analyses and to ensure rigor of the results. In this study we compare state of the art Deep Learning architectures for the use case of multispectral image classification on datasets from three different domains to determine whether there is a suitable architecture for all applications or if a domain-specific architecture is required. Experiments have shown that there are multiple Convolutional Neural Network (CNN) architectures that show comparable results with regard to the evaluation criteria accuracy and computational effort. A single architecture that outperforms other architectures in all three domains could not be found.**

***Index terms*—Computer Vision, Classification, Deep Learning, Multispectral Imaging Flow Cytometry**

I. Introduction

MULTISPECTRAL imaging flow cytometry (MIFC) has been recently shown to be useful for environmental monitoring of plant-pollinator interactions and assessment of food and water quality [1], [2], [3]. This measuring technique, originally designed for immunological analyses as blood cell analysis, allows to separate single cells from a fluid suspension by hydro-dynamically focusing of a sample stream in a narrowing flow cell, surrounded by a sheath stream. The resulting acceleration of the cells, in conjunction with the hydrodynamic forces acting on them, separates the cells in the liquid flow. Once the cells occur as single cells in the sample stream they will be imaged by CCD (charge-coupled device) cameras. Two cameras take independent images from LED illumination of the cells (brightfield images), images from different laser excitation and respective light emission of the cells (fluorescence images) as well as images from red light illumination (scatter images). The instrument is capable of capturing 12 images per particle (two brightfield, nine fluorescence and one scatter image) with a sample throughput of up to 2000 particles/s [1], [2], [4]. The high sample throughput allows for an unprecedentedly high measuring efficiency in comparison to the traditional benchmark of manual microscopy. The images produced by the imaging flow cytometer instrument come with some specifics in comparison to other images used in common image recognition tasks. As the microscopic images are recorded in high throughput, they have a relatively low resolution (a pixel size of 0.5 × 0.5 µm) but at the same time, a high number of images per sample can be collected (500-5000 particles à 12 images/particle). In addition, MIFC already provides images where the object of interest is depicted in the image center and the images have a uniform background. We could already demonstrate the successful application of the different architectures ResNet V2 [4] and Inception V3 [2] for this kind of data to distinguish 27 classes of phytoplankton (relevant for water quality assessment) and 35 classes of pollen (relevant for plant-pollinator studies, food and air quality), respectively. [5] used a ResNet50 architecture to discriminate seven classes of different blood cell morphologies. These examples already show that for different MIFC datasets, different architectures have been applied. But no systematic evaluation of different architectures has taken place so far. The different mentioned application examples have different numbers of classes which need to be differentiated. For immunological applications, a limited set of cell types needs to be distinguished, while in environmental monitoring, potentially several thousands of classes need to be differentiated eventually [3].

Most models used to classify MIFC datasets are based on architectures that were assessed on large image datasets such as ImageNet [1], [2]. The images of the pre-trained dataset differ from MIFC images in the way that they depict

We thank the German Research Foundation for support via the iDiv Flexpool Funding projects Grant Number: Grant Number: 34600830-13, ('PolDiv'-Project) and Grant Number: RA-373/20 (iCyt - Support Unit). Furthermore, we thank Prof. Dr. Patrick Maeder and David Boho from University Ilmenau for their valuable thoughts during the preparation of the study and all student helpers and interns for help during image annotation.

©2023, PTI

Thematic track: Data Science in Health, Ecology and Commerce

RGB images, have more labels and higher resolution. To acknowledge the specificity of the images and the multiple channels available in MIFC datasets, we want to enhance the existing studies and aim to answer the following research questions:

- What is the best architecture for different kinds of datasets with respect to accuracy?
- What is the most suitable degree of CNN architecture complexity for the individual datasets?
- From the best performing architectures which is the most sustainable one with respect to computational effort and resource consumption?
- Is there a general architecture for all different MIFC datasets with best performance on accuracy and computational effort?

The remainder of the paper is structured as follows: Section II presents the applied methods to answer the research questions. Section III presents the obtained classification results. A discussion of the results obtained with reference to the initial research questions, takes place in section IV. Possible limitations of the work are addressed in Section V. The study concludes with a summary of all findings and an outlook on future work in section VI.

II. Methods

The methodology used, from the selection of CNN architectures, data acquisition, data preparation, MIFC measurements to training and validation strategy used, is described below.

Selection of CNN architectures

For the comparison of different CNN architectures we considered seven different architectures: DenseNet [6], Inception V3 [7], Inception-ResNet V2 [8], MobileNetV2 [9], ResNet V2 [10], VGG [11] and Xception [12]. DenseNet (121, 169, 201), ResNet (50, 101, 152) and VGG (16, 19) have different depth configurations that were also considered. These architectures represent frequently used architectures in science and practice and thus were selected for this architecture comparison. Based on this architecture selection, it is also possible to assess whether more complex architectures are more suitable than shallower architectures for classification, and what correlation there is between the complexity of an architecture and resource consumption during model training. Table I provides an overview of the selected architectures with the specific network depths and input size.

TABLE I: Overview of CNN architectures

Model	Depth	Input Size
DenseNet-121	121	224x224x3
DenseNet-169	169	224x224x3
DenseNet-201	201	224x224x3
Inception V3	48	229x229x3
Inception-ResNet V2	164	229x229x3
MobileNet V2	53	224x224x3
ResNet-50	50	224x224x3
ResNet-101	101	224x224x3
ResNet-152	152	224x224x3
VGG-16	16	224x224x3
VGG-19	19	224x224x3
Xception	71	229x229x3

Dataset acquisition and annotation

The availability of high-quality datasets is crucial for the performance of a supervised learning approach. To ensure the cross-domain applicability of our study, we used three datasets from different fields of application. The stored red blood cells (RBC) dataset consists of 63,000 samples in seven clinically relevant blood cell morphologies associated with storage lesions [5].

The phytoplankton dataset consists of six naturally co-occurring species, three cyanobacteria (*Chroococcus minutus* - species "C" SAG 41.79, *Microcystis aeruginosa* - species "M" SAG 1450-1, *Synechocystis sp.* - species "S" PCC 6803, and three chlorophytes *Acutodesmus obliquus* - species "A" SAG 276-3a, *Desmodesmus armatus* - species "D" SAG 276-4d, and *Oocystis marssonii* - species "O" SAG 257-1), grown exponentially under controlled laboratory conditions (60 µmol photons/ m-2/s-1; 14/10 h light/dark cycle; WC-Medium) [13]. Measurements were performed in subsequent biologically independent mono-culture experiments (rep 0, rep 1). The total dataset contains 12,000 samples of species A,C,D,M,O and S.

The pollen dataset samples were collected in the botanical garden of Leipzig, Germany and natural fields in the surrounding area of Leipzig during peak flowering time from 2018 to 2020. It consists of 4,800 samples, which are randomly stratified, i.e. each class has the same number of samples. There are twelve species in seven genera, which are mostly wind-distributed and comprise morphological similar allergenic species. Similarity in pollen features within a genus restricts classical light microscopic discrimination of the selected species to the genus level. For that reason we tested classification on different taxonomic levels (genus and species) to see if CNN were even capable of classifying on species level. To get a representative dataset, we used 20% high-quality images (pollen in focus, non-cropped, without other debris particles on image) and 80% low-quality images (pollen either out of focus, partially cropped or pollen images with additional debris particles on image).

MIFC measurements

All samples were measured with an Amnis® Image Stream®X MK II imaging flow cytometer (Amnis part of Cytek, Amsterdam, Netherlands). For the phytoplankton dataset, measurements were performed according to [13], for the blood cell dataset according to [5] and for the pollen dataset according to [2] as shown in Table II.

Data preparation and augmentation

All images were channel-wise standardized (i. e. rescaled to have a mean of 0 and unit variance) and normalized (i. e. rescaled to a value range of 0 and 1) utilizing the pixel values from the respective training dataset.

Most CNN architectures are translation invariant but not invariant with regard to scale, rotation and different perturbations. To address the problem of overfitting we artificially

TABLE II: Datasets

Characteristics	Phytoplankton	Pollen	Blood cell
Classes	6	12	7
No. particles	12000	4800	63000
Channels	12 (1-12)	7 (1-6,9)	3 (1,9,12)
Lasers	3 (488/561/785)	3 (488/561/785)	5
Magnification	40x	40x	60x
Sheath fluid	D-PBS	D-PBS	PBS
References	[13]	unpublished	[5]

increased the dataset size to alleviate scarcity issues. Several random data augmentations that yield credible images are introduced to the training datasets to help the models to generalize better and to be more robust with regard to random perturbations and noise [14].

Brightness and contrast are randomly adjusted channel-wise in [-0.3, 0.3] and [0.5, 2.0] intervals to achieve a robustness of the classifier for different MIFC calibrations, fluorescence and random background noise. Further random geometric transformations as rotation, horizontal flipping and central cropping (interval [0.8, 1.0]) are introduced to make the classifier robust against different cell orientations and cell sizes that may occur across different measurements. As all images have varying aspect ratios they were resized to 116 by 116 pixels, padded with zeros.

Training and validation strategy

For each dataset a k-fold stratified cross-validation [15] was performed to find an optimal hyperparameter combination that is less biased or optimistic compared to a simple train/validation/test split. For that purpose the datasets were split into k=5 equally-sized subsets that have the same class distribution as the original datasets. For each subset the set is used as a test dataset on which the model performance is evaluated and the remaining sets are used to train the model. Over all runs the performance metrics accuracy, macro averaged precision, recall and F_1 score were averaged to get an estimate how good the final model performs and how robust it is with regard to data variability.

All models were trained on all available multispectral channels (1,9,12 for RBC, 1-12 for phytoplankton and 1-6 and 9 for wind pollen).

A grid search was used for hyperparameter optimization. Considered hyperparameters were optimizer function (RMSProp [16], Adam [17]), batch size (16, 32, 64) and learning rate (1e-4, 1e-5, 1e-6). For all architectures categorical cross entropy was chosen as a loss function. In total, 18 hyperparameter combinations were evaluated per architecture.

The learning rate was reduced by a factor of 10 if the validation loss had not decreased within 20 epochs and the training was stopped when the validation loss had not decreased within 30 consecutive epochs.

The best hyperparameter combination for each architecture was assessed on a fixed holdout dataset eventually. For each model the number of trainable parameters (weights and biases) was calculated as a measure of model complexity. As the number of available image channels varies between the datasets, the number of parameters for the same architecture differs. Additionally we measured the floating point operations (FLOPs) for a single forward pass to quantify inference performance.

III. Results

We wanted to find a CNN architecture that shows the best metrics for MIFC datasets from different domains and determine whether complex architectures outperform simpler ones.

Three datasets recorded with an imaging flow cytometer containing samples from three different application domains, i.e., wind pollen, phytoplankton and blood cells, were used to evaluate the different CNN architectures captured in Table I. These CNN architectures were trained on the different datasets. The task of each model was to recognize patterns and structures in the images in order to achieve the highest possible accuracy in assigning the images to their respective classes.

Classification Results

Table III and Table IV show the results for each architecture on the species or on the genus level. The balanced wind pollen dataset shows an increase of performance metrics from the training on the species level to the training on the same dataset on genus level. This is not surprising, as the number of classes are reduced from twelve to seven and the number of samples per class is increased which gives the classifier during training more exposure to training samples and thus the potential to generalize better. We observed that the Inception-ResNet is the best performing architecture (96.88% accuracy on species level, 98.96% on genus level) in both tests, closely followed by Inception, DenseNet and Xception. An impact of the size and complexity of the model on the classification accuracy cannot be determined on the basis of the results obtained.

The classification results for the phytoplankton dataset are shown in Table V (train on rep-0, test in rep-1) and Table VI (train on rep-1, test in rep-0). The phytoplankton dataset was trained on one independent replicate of the measurement, evaluated on another independent replicate of the measurement and vice versa. In both tests the VGG-16 and VGG-19 architecture showed the best classification accuracy of approximately 92%, closely followed by Inception and DenseNet architectures with comparable accuracies. Again the size of the models regarding number of parameters seems not to have a real impact on the accuracy of the model.

Table VII (train on Canadian, test on Swiss) and Table VIII (train on Swiss, test on Canadian) show the results for the blood quality dataset. We observed that there is no single prevailing architecture for the blood cell dataset that was trained on samples originating from Switzerland, evaluated on samples originating from Canada and vice versa. DenseNet-121 and Xception are both the best performing architectures. Here we observe F_1 scores of 75.68% (train on Canadian, test on Swiss) and 87.90% (train on Swiss, test on Canadian).

TABLE III: Wind pollen (species level)

Architecture	Top-1 Acc.	Precision	Recall	F_1 score
Inception-ResNet	96.88%	96.89%	96.88%	96.87%
Inception	96.25%	96.29%	96.56%	96.24%
DenseNet-121	96.25%	96.33%	96.46%	96.24%
Xception	96.04%	96.09%	96.35%	96.04%
DenseNet-169	96.04%	96.19%	96.29%	96.01%
ResNet-152	95.83%	95.95%	96.22%	95.81%
ResNet-50	95.63%	95.72%	96.13%	95.60%
DenseNet-201	95.42%	95.47%	96.04%	95.43%
ResNet-101	95.42%	95.80%	95.97%	95.43%
VGG16	94.38%	94.57%	95.66%	94.39%
VGG19	94.38%	94.42%	95.56%	94.37%
MobileNet	89.38%	89.52%	94.84%	89.34%

TABLE IV: Wind pollen (genus level)

Architecture	Top-1 Acc.	Precision	Recall	F_1 score
Inception-ResNet	98.96%	98.95%	98.81%	98.87%
Inception	98.75%	98.82%	98.69%	98.69%
DenseNet-201	98.75%	98.42%	98.59%	98.64%
DenseNet-169	98.75%	98.88%	98.66%	98.62%
Xception	98.54%	98.77%	98.58%	98.51%
VGG16	98.54%	98.71%	98.53%	98.48%
DenseNet-121	98.33%	98.28%	98.44%	98.08%
ResNet-152	98.33%	98.05%	98.40%	98.07%
ResNet-101	97.71%	97.54%	98.34%	97.69%
VGG19	97.71%	97.36%	98.30%	97.63%
ResNet-50	97.50%	97.40%	98.06%	97.23%
MobileNet	96.46%	96.32%	97.67%	96.14%

TABLE V: Phytoplankton (train: rep-0, test: rep-1)

Architecture	Top-1 Acc.	Precision	Recall	F_1 score
VGG16	92.62%	92.72%	92.62%	92.54%
VGG19	92.50%	92.54%	92.56%	92.46%
Inception	92.27%	92.23%	92.46%	92.20%
Inception-ResNet	92.03%	92.03%	92.35%	91.97%
DenseNet-121	91.52%	91.49%	92.19%	91.45%
DenseNet-169	91.47%	91.45%	92.07%	91.41%
ResNet-101	91.22%	91.19%	91.95%	91.16%
DenseNet-201	91.20%	91.15%	91.85%	91.15%
Xception	91.18%	91.14%	91.78%	91.12%
ResNet-152	90.53%	90.47%	91.65%	90.45%
ResNet-50	90.37%	90.27%	91.45%	90.26%
MobileNet	87.32%	87.12%	90.69%	87.16%

TABLE VI: Phytoplankton (train: rep-1, test: rep-0)

Architecture	Top-1 Acc.	Precision	Recall	F_1 score
VGG16	91.88%	92.01%	91.88%	91.83%
VGG19	91.87%	92.10%	91.87%	91.82%
DenseNet-121	91.87%	91.97%	91.87%	91.82%
DenseNet-201	91.38%	91.64%	91.75%	91.35%
Inception	91.30%	91.52%	91.66%	91.27%
ResNet-152	91.28%	91.43%	91.60%	91.23%
DenseNet-169	91.17%	91.32%	91.54%	91.11%
ResNet-101	90.92%	91.03%	91.46%	90.87%
Xception	90.42%	90.61%	91.34%	90.38%
ResNet-50	90.28%	90.45%	91.24%	90.20%
Inception-ResNet	90.10%	90.46%	91.13%	90.12%
MobileNet	87.45%	87.86%	90.39%	87.27%

TABLE VII: Blood quality (train: Canadian, test: Swiss

Architecture	Top-1 Acc.	Precision	Recall	F_1 score
DenseNet-121	87.09%	75.07%	81.47%	75.68%
DenseNet-201	86.75%	74.63%	80.53%	75.25%
Inception-ResNet	86.48%	76.29%	79.58%	76.42%
DenseNet-169	86.40%	74.62%	81.02%	75.28%
Inception	86.27%	75.54%	80.14%	76.04%
ResNet-152	85.96%	74.98%	80.58%	75.14%
VGG16	85.93%	73.76%	79.23%	74.23%
ResNet-50	85.69%	72.34%	80.29%	73.47%
ResNet-101	85.35%	72.65%	79.97%	73.11%
VGG19	85.33%	73.85%	81.06%	75.09%
Xception	84.85%	76.07%	79.73%	75.65%
MobileNet	83.89%	71.99%	78.56%	72.91%

TABLE VIII: Blood quality (train: Swiss, test: Canadian)

Architecture	Top-1 Acc.	Precision	Recall	F_1 score
Xception	87.91%	88.78%	87.71%	87.90%
Inception	87.90%	88.68%	87.67%	87.83%
ResNet-152	87.84%	88.79%	87.91%	87.95%
Inception-ResNet	87.79%	88.54%	87.84%	87.79%
DenseNet-121	87.71%	88.67%	87.82%	87.86%
VGG19	87.68%	88.48%	87.99%	87.75%
DenseNet-201	87.55%	88.68%	87.62%	87.64%
ResNet-50	87.26%	88.48%	87.31%	87.41%
VGG16	87.12%	88.22%	87.35%	87.29%
DenseNet-169	87.04%	88.25%	87.25%	87.23%
ResNet-101	86.85%	88.13%	86.91%	86.97%
MobileNet	86.63%	87.86%	86.41%	86.60%

There is a noticeable difference in the F1 scores obtained between train on Canadian and test on Swiss and vice versa.

The ability of the models to distinguish between classes of each dataset is illustrated in Fig. 1 using the associated confusion matrix. In (a), the accuracies of the mappings for wind pollen at species level are given. The Inception-ResNet achieves optimal assignments for the vast majority of pollen classes and shows only slight weaknesses in distinguishing pollen from the same genus level (e.g., *Corylus avellana* and *Corylus colurna*). These pollen species from the same genus level show many similarities in their appearance, so that misclassifications can occur here. For the genus-level class assignment in (b), it can be seen that the accuracy of the Inception-ResNet could be increased compared to the species-level assignment. Compared to (a), no discrimination on species level is necessary, allowing the model to more easily identify differences in classes at the genus level. Confusion matrix (c) shows the accuracy of the VGG16 network class assignments for phytoplankton (train on rep-0, test on rep-1), (d) for phytoplankton (train on rep-1, test on rep-0). No significant differences can be detected between the two variants in terms of accuracy in class assignment. It is noticeable that the model shows similar weaknesses in correct assignment for classes 'C' (*Chroococcus minutus*) and 'O' (*Oocystis marssonii*) for both variants. The shape and size of both species is quite similar which might explain a low discriminatory power. Since both species belong to different taxonomic groups, the discrimination could be enabled with flow cytometric taxonomic separation [18]. The accuracy of class assignment for the Blood Quality dataset is shown in confusion matrix (e) (train on Canadian, test on Swiss) and (f) (train on Swiss, test on Canadian). Again, a similar level of assignment accuracy to classes can be observed for both variants. In (e), however, the class CrenatedSpheroid with an accuracy of only 52.3% shows a conspicuously high number of false assignments, which cannot be observed to the same

extent in (d). In general, the blood cell group assignment is more complicated than species identification in the two other datasets and for a distinction there is less a discrete difference but rather a continuously diverging morphology.

FLOPs - Floating Point Operations

In addition to identifying a suitable architecture for the highest possible classification accuracy, the computational effort and associated resource consumption of the architectures was considered in terms of FLOPs, number of floating point operations. Therefore, top-1 accuracy was set in relation to the number of FLOPs (in billions) and the number of model parameters (in millions). Fig. 2 (a) and (b) illustrates this comparison across the model architectures for the wind pollen dataset. With regard to the classification of wind pollen at the species level, it was observed that the Top-1 Accuracy decreased with increasing number of FLOPs. For classification on genus level this tendency is not observed. On both levels the best Accuracy-Flops ratio, and thus most resource efficient architecture, is provided by the Inception-ResNet architecture. Fig. 2 (c) and (d) illustrates this comparison for the phytoplankton dataset. For (c) Phytoplankton (train on rep-0, test on rep-1) the best Accuracy/FLOP ratio is achieved by the Inception network. Here, a relativly high accuracy can be achieved on a small number of FLOPs. That means that less operations are required to run a single instance of the Inception model compared to VGG16 or VGG19 to achieve similar accuracy. The same can be stated for (d) Phytoplankton (train on rep-1, test on rep-0). Here, the DenseNet-121 architecture achieves the best Accuracy/FLOP ratio and allows a resource-efficient use of the model. For the blood quality dataset. shown in Fig. 2 (e) and (f), DenseNet-169 achieves the best ratio for (e) RBC (train on Canadian, test on Swiss) and the Inception networks was found to be the most ressource-efficient one for (f) RBC (train on Swiss, test on Canadian).

IV. Discussion

In the present work, different architectures of artificial neural networks are analyzed with respect to their performance for the classification of different datasets (wind pollen, phytoplankton, blood cells) generated by MIFC. With reference to the research questions raised, the following findings could be obtained based on this study:

1) What is the best architecture for different kinds of datasets with respect to accuracy? Our findings differ from applied techniques used in previous literature, where different network architectures were used to classify pollen, phytoplankton and blood cells. In [4], the authors achieved an average accuracy of 99% for combined images of phytoplankton at species level using a ResNet v2 with 50 convolutional layers, which was not surpassed by the VGG16 as the best performing model used in our analysis. In this context, it should be noted that the models used here are not precisely tuned, since the focus of this work is on the comparison of different architectures than on the optimization of a single architecture.

For the classification of wind pollen, the authors in [2] achieve an accuracy of max. 96% with an Inception V3 network with 48 convolutional layer. These results can be confirmed within the scope of this study, so that similar results could be achieved with the compared architectures (e.g., Inception-ResNet (164 layers) = 96.88%, Inveption V3 (48 layers) = 96.25%). This suggests, that deeper networks are not necessarily superior to shallower networks for pollen classification.

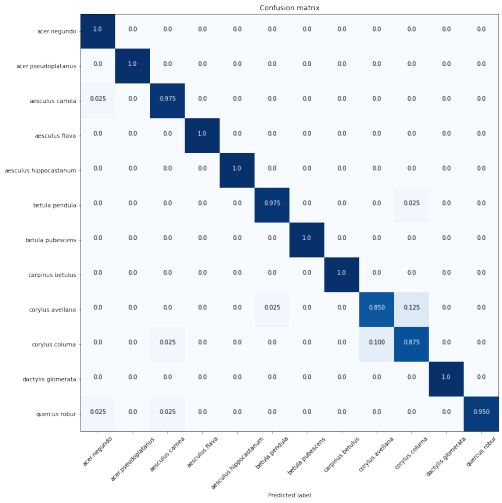
With respect to the classification of the blood cell data set, it can be highlighted that both similarities and differences can be identified in recent studies. The authors in [5] used a ResNet-50 for a classification of red blood cells. The results show that a comparably good classification accuracy could be achieved and that similar difficulties exist in the correct assignment of certain classes (e.g. crenated spheroid). The ResNet-50 in the comparative study achieves an average accuracy of 80%. The best performing architectures in this study are DenseNet-121 (87.09% for train on Canadian, test on Swiss) and XCeption (87.91% for train on Swiss, test on Canadian).

2) What is the most suitable degree of CNN architecture complexity for the individual datasets? We could show that deeper neural networks do not necessarily perform better than shallow networks. Instead, an accurate classification may be achieved with comparably shallow networks, such as VGG-16, VGG-19 or Inception (48 layers). This fact leads to the conclusion that the use of such, shallower networks would be advantageous, especially in the case of limited hardware resources.

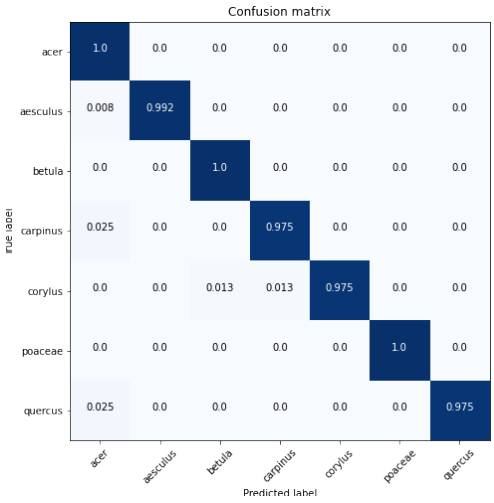
3) From the best performing architectures which is the most sustainable one with respect to computational effort and resource consumption? The advantage of using CNN with fewer layers is that they have lower hardware requirements and shorter training times compared to their deeper counterparts. Shorter training times allow testing of more hyperparameters and simplify the overall training process. This is particularly useful in environments with limited resources or where a resource-efficient use deep learning techniques is aspired.

Additionally, shorter training times can facilitate the integration of improvement methods into the training data, such as the implementation of "human in the loop" annotations. Human in the loop means that the training of a network is monitored by a human expert who can intervene at critical steps and correct the network. For example, the expert can check misclassifications, effectively reducing annotation noise. With shorter training times, such feedback loops can be executed more quickly.

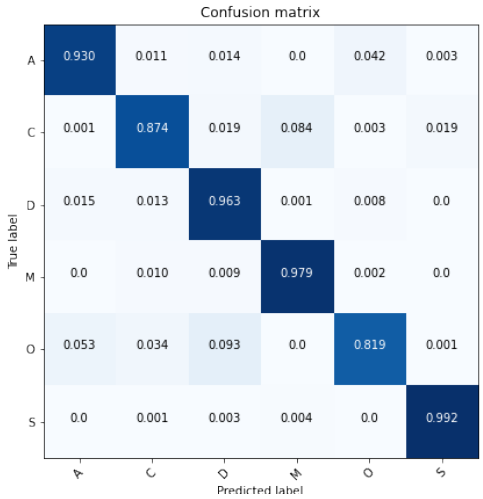
4) Is there a general architecture for all different MIFC datasets with best performance on accuracy and computational effort? Overall, it can be highlighted that no single best architecture could be identified for the respective datasets, as they are often very close in terms of accuracy (deviations in many cases under 1%). It can be emphasized that there is no best-performing architecture from a generally valid point of view with regard to the accuracy-resource ratio.



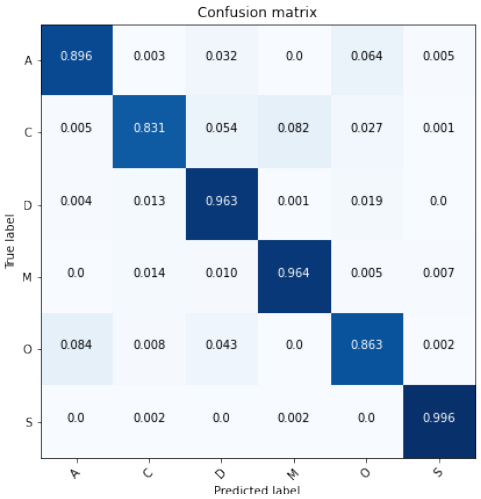
(a) Confusion matrix Inception-ResNet for wind pollen (species level)



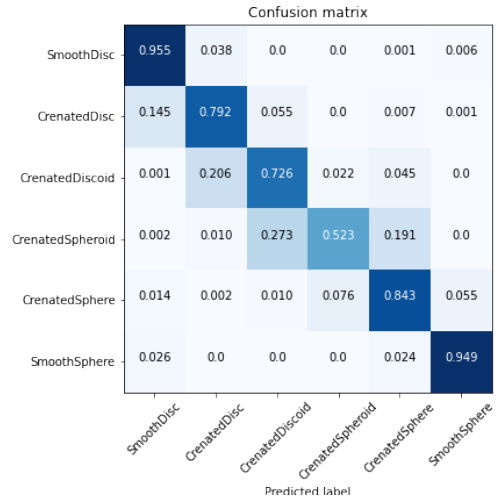
(b) Confusion matrix Inception-ResNet for wind pollen (genus level)



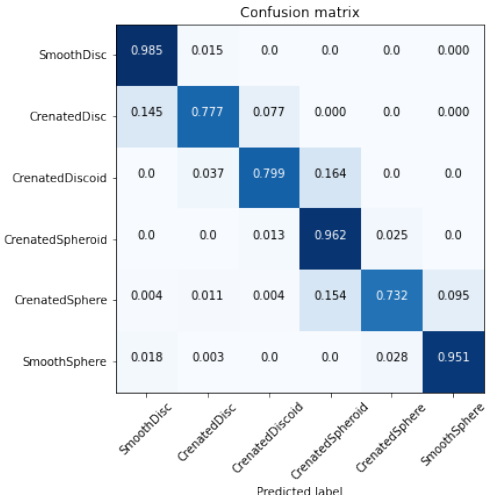
(c) Confusion matrix for VGG16 on phytoplankton (train on rep-0, test on rep-1)



(d) Confusion matrix for VGG16 on phytoplankton (train on rep-1, test on rep-0)

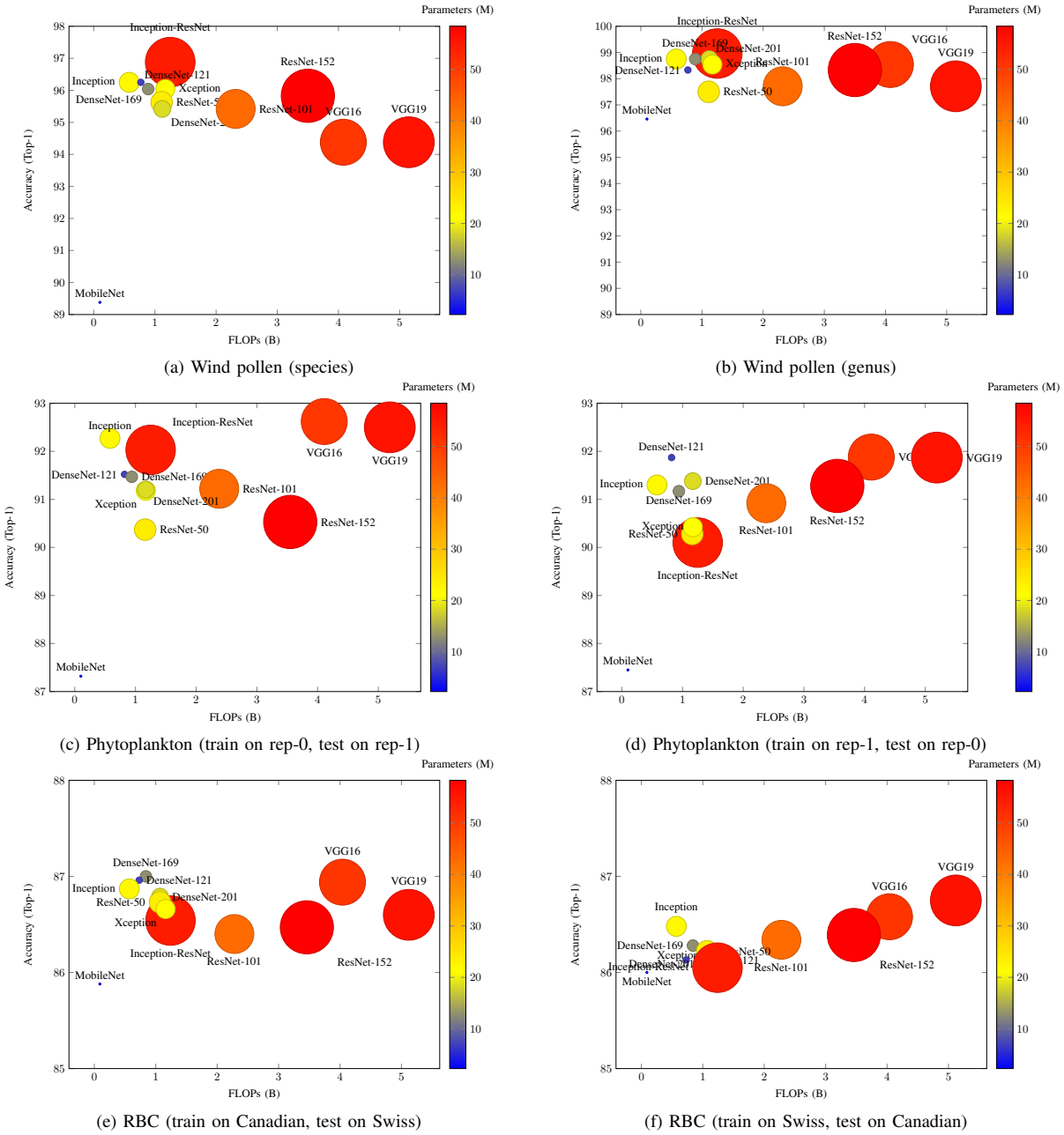


(e) Confusion matrix for DenseNet-169 on blood quality (train on Canadian, test on Swiss)



(f) Confusion matrix for VGG19 on blood quality (train on Swiss, test on Canadian)

Fig. 1: Confusion matrices of the best performing models on the respective datasets



(a) Wind pollen (species)

(b) Wind pollen (genus)

(c) Phytoplankton (train on rep-0, test on rep-1)

(d) Phytoplankton (train on rep-1, test on rep-0)

(e) RBC (train on Canadian, test on Swiss)

(f) RBC (train on Swiss, test on Canadian)

Fig. 2: FLOPs (Billions) compared to Accuracy (Top-1)

Depending on the data set, different architectures achieve an optimal accuracy-resource ratio.

V. Limitations

A limitation of the present work is, that different, mostly larger architectures, may tend to overfit on less complex datasets, like the blood cell dataset, that is less complex than the wind pollen dataset. As a consequence, an overfitting with a lower generalizability of some results cannot be excluded and should be considered when interpreting our results. Furthermore, the CNN architectures used in this study not have fully optimized hyperparameters, which could impact the performance of the models. Those architectures may not performed as well as they could have following hyperparameter optimization.

VI. Conclusions

The aim of the present work was to compare different CNN architectures for the classification of MIFC datasets. In this context, seven different architectures with different complexity and depth were trained and tested on three datasets

(wind pollen, phytoplankton, red blood cells). The evaluation results demonstrate, that complex architectures with a large number of trainable parameters or increasing depth are not always required to achieve state of the art results. A DenseNet architecture with 121 layers reaches comparable results to more complex architectures such as Inception-ResNet and VGG that consume significantly more computing resources during training. The ecological footprint during model training and inference can be reduced by using simpler architectures without sacrificing accuracy.

A CNN architecture that is most qualified for all datasets under consideration could not be determined, as most considered architectures show comparable results with regard to the evaluation criteria.

Future research could include training and testing on even larger datasets with more classes and higher variability. In addition, hyperparameter optimizations can be performed on individual architectures to identify a universally best architecture for the classification of the investigated datasets. Furthermore, it has to be evaluated whether the use of a single architecture is reasonable at all and can be complemented by the implementation of ensemble methods.

ACKNOWLEDGMENT

Computations for this work were done in part using resources of the Leipzig University Computing Center.

AVAILABILITY OF DATA AND MATERIALS

Data and materials used in the study are available upon reasonable request.

REFERENCES

[1] S. Dunker, "Hidden Secrets Behind Dots: Improved Phytoplankton Taxonomic Resolution Using High-Throughput Imaging Flow Cytometry," *Cytometry Part A*, vol. 95, no. 8, pp. 854–868, 2019, doi: 10.1002/cyto.a.23870.

[2] S. Dunker, E. Motivans, D. Rakosy, D. Boho, P. Mäder, T. Hornick, and T. M. Knight, "Pollen analysis using multispectral imaging flow cytometry and deep learning," *New Phytologist*, vol. 229, no. 1, pp. 593–606, 2021. doi: 10.1111/nph.16882

[3] S. Dunker, M. Boyd, W. Durka, S. Erler, W. S. Harpole, S. Henning, U. Herzschuh, T. Hornick, T. Knight, S. Lips *et al.*, "The potential of multispectral imaging flow cytometry for environmental monitoring," *Cytometry Part A*, vol. 101, no. 9, pp. 782–799, 2022. doi: 10.1002/cyto.a.24658

[4] S. Dunker, D. Boho, J. Wäldchen, and P. Mäder, "Combining high-throughput imaging flow cytometry and deep learning for efficient species and life-cycle stage identification of phytoplankton," *BMC Ecology*, vol. 18, no. 1, pp. 1–15, 2018. doi: 10.1186/s12898-018-0209-5

[5] M. Doan, J. A. Sebastian, J. C. Caicedo, S. Siegert, A. Roch, T. R. Turner, O. Mykhailova, R. N. Pinto, C. McQuin, A. Goodman, M. J. Parsons, O. Wolkenhauer, H. Hennig, S. Singh, A. Wilson, J. P. Acker, P. Rees, M. C. Kolios, A. E. Carpenter, and D. Geman, "Objective assessment of stored blood quality by deep learning," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 117, no. 35, pp. 21 381–21 390, sep 2020. doi: 10.1073/pnas.2001227117

[6] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, vol. 2017-January, pp. 2261–2269, 2017. doi: 10.1109/CVPR.2017.243

[7] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the Inception Architecture for Computer Vision," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2016-Decem, pp. 2818–2826, 2016. doi: 10.1109/CVPR.2016.308

[8] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, "Inception-v4, inception-ResNet and the impact of residual connections on learning," in *31st AAAI Conference on Artificial Intelligence, AAAI 2017*. AAAI press, feb 2017. doi: 10.48550/arXiv.1602.07261 pp. 4278–4284.

[9] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L. C. Chen, "MobileNetV2: Inverted Residuals and Linear Bottlenecks," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 4510–4520, 2018. doi: 10.1109/CVPR.2018.00474

[10] K. He, X. Zhang, S. Ren, and J. Sun, "Identity Mappings in Deep Residual Networks," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 9908 LNCS, pp. 630–645, mar 2016. doi: 10.1007/978-3-319-46493-038

[11] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," sep 2015. doi: 10.48550/arXiv.1409.1556

[12] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, vol. 2017-January. Institute of Electrical and Electronics Engineers Inc., nov 2017. doi: 10.1109/CVPR.2017.195. ISBN 9781538604571 pp. 1800–1807.

[13] P. Hofmann, A. Chatzinotas, W. S. Harpole, and S. Dunker, "Temperature and stoichiometric dependence of phytoplankton traits," *Ecology*, vol. 100, no. 12, p. e02875, dec 2019. doi: 10.1002/ecy.2875

[14] C. Shorten and T. M. Khoshgoftaar, "A survey on Image Data Augmentation for Deep Learning," *Journal of Big Data*, vol. 6, no. 1, 2019. doi: 10.1186/s40537-019-0197-0

[15] X. Zeng, T. R. Martinez, X. Inchuan Zeng, and T. N. R M A Rtinez, "Distribution-balanced stratified cross-validation for accuracy estimation," *Journal of Experimental & Theoretical Artificial Intelligence*, vol. 12, no. 1, pp. 1–12, 2000. doi: 10.1080/095281300146272

[16] S. Ruder, "An overview of gradient descent optimization algorithms," 2017. doi: 10.48550/arXiv.1609.04747

[17] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2017. doi: 10.48550/arXiv.1412.6980

[18] S. Dunker, "Hidden secrets behind dots: Improved phytoplankton taxonomic resolution using high-throughput imaging flow cytometry," *Cytometry Part A*, vol. 95, no. 8, pp. 854–868, 2019. doi: 10.1002/cyto.a.23870

Position Papers of the 18th Conference on Computer Science and Intelligence Systems pp. 67–76
DOI: 10.15439/2023F5882
ISSN 2300-5963 ACSIS, Vol. 36

Decoupling Types and Representations of Values for Runtime Optimizations*

Krajča Petr† and Škrabal Radomír
0000-0003-4278-3130
0000-0003-3929-7238
Palacky University Olomouc
17. listopadu 12, CZ77146 Olomouc
Czech Republic
E-mail petr.krajca@upol.cz, radomir.skrabal01@upol.cz

***Abstract*—The way data is stored in a computer's memory is crucial from the performance point of view. The choice of the most appropriate data structure depends not only on the algorithm but also on the input data itself. Unfortunately, this information may not always be known in advance, requiring programmers to make educated guesses about the data's characteristics. If these guesses are inaccurate, it can result in suboptimal performance. To address this challenge, we introduce a novel programming language that draws inspiration from database management systems. This language has the capability to automatically select optimal data structures and, consequently, algorithms based on the input data to improve performance.**

I. Introduction

DATA types play an increasingly important role in modern programming. In fact, it is not a single role, but four roles at once. Data types (i) create abstractions, (ii) define variable domains (which values can variable hold), (iii) define representation of values (how data are stored in memory), and corollary (iv) assists to dispatch operations with data [1].

Much focus has been given to the role (ii) of data types which allows to build a more robust software, hence simplify software development. We are to explore the role (iii) which deals with value representation, since it has significant impact on performance, due to the role (iv) of the type system.

Let us recall what is meant by type, value, and its representation. For simplicity we are to assume that *a type* is a set of values. *A value* can be seen as an abstract entity which has no location in time or space. However, a value may have encoding that can be represented in the memory of the computer. This encoding of a value shall be called *representation* and can be assigned to a variable. Naturally, a single value may have multiple representations. For instance *value 42* may be represented as "42" (two numerals), "forty two" (words), "XLII" (roman numerals), 00101010 (binary number), etc. All of these are valid representations of the number, however, each is suitable for different situations. Note that the notions of a *type* and *representation* are orthogonal. The type of a variable provides information on values that can be assigned to a given variable, while representation is an encoding of a value, see [2]. Types play an important role during the compilation process, since they allow to discover invalid operations, representations are crucial for program execution since they determine how the program is to be executed. For example, points can be represented in Cartesian or polar coordinates while each representation is suitable for different applications. Analogously, in data analysis various representations of vectors (dense or sparse) and matrices (dense, sparse, row/column major) are used. Which of these representations is more appropriate is determined by the algorithm and often by the input data itself.

In contemporary programming languages decoupling of types and their representations is achieved via interface inheritance (Java, C#) or type-classes (Haskell). These time-approved techniques have a particular downside. The selection of a type representation is often made in compile-time based on more-or-less accurate assumptions on the future use of the program. If the data representation does not fit the actual input data, it may lead to a suboptimal performance, e.g., if a sparse matrix is stored as a dense matrix in memory.

To deal with this kind of issues we propose a programming language where types and representations of values are strictly decoupled. Our motivation comes from the world of relational database management systems (RDBMS). For instance, RDBMS stores tuples in a sequential data file. Depending on the size of data, meta-data and query, RDBMS during query processing selects the most suitable representation of the data from the data file. Let us consider join of two data tables in RDBMS. If the data are small enough then the input is read and processed sequentially using some trivial algorithm (e.g. nested loop join). If the data are larger, RBDMS can change their representation to process it faster. For instance, to a hash table and process data using a hash-join. Or, if an index is available, data can be turned into an ordered set and processed using a sort-merge-join algorithm, see e.g. [3] for details.

Our intention is to design a language where this capability, i.e., select the most appropriate data representation and by extension most appropriate algorithms, is available in a general purpose programming language. The intended applications come from the area of data analysis where the selection of an appropriate data representation is of a crucial importance.

*Supported by grants IGA_PrF_2022_018 and IGA_PrF_2023_026 of Palacký University Olomouc.
†Corresponding author.

The paper is organized as follows. Section II provides introduction to the basic of the language we propose. Then, Section III describes the concept of an *extended function* allowing to dispatch functions based on types and values of arguments. Sections IV and V are devoted to the type inference in our language. Paper concludes with Sections VI and VII on the experimental evaluation and notes on the current state of the language in the future.

II. Language Velka

We propose a programming language called Velka as an experimental framework, where types and representations are strictly separated from each other. Velka is loosely based on Scheme [4] and has a Lisp-like syntax.

Velka is strongly typed, detecting many type-related errors at compilation time, with no automatic type coercion. However, Velka can convert between multiple representations of a single type if it is viable. The evaluation method (strict or lazy) for Velka is not specified.

The following subsections are brief introduction to Velka language. We tacitly assume that readers are acquitted with the Lisp-like language [4], [5], [6], therefore we focus on the most essential parts of the language.

A. Type and representation signatures

Representation signature is a way of expressing a representation statically in Velka syntax. Atomic types use their name as a signature like `Int`, `Bool` for integers or Boolean values. An atomic representations uses colon `:` to separate the type and its representation, for example `Int:Native` and `Int:Roman` denote two different representations of integers, (i) a binary representation in the memory and (ii) a string of Roman numerals. Type signatures of tuples are written in parentheses, e.g., `(Int Bool)`.

B. Expressions

Velka follows in the majority of cases syntax and semantics of the Lisp-family languages. Expressions are either atoms (literals, symbols) that evaluate to their associated values or lists, e.g., `(+ 1 2)` or `(+ 1 (* 2 3))`. Lists are evaluated recursively as follows. First, the first element of the list is evaluated. If it evaluates to a function, then all remaining elements (i.e., arguments) are evaluated, and the function is applied on these values. Besides functions, the first element may evaluate to a special form (`if`, `define`) that is a special operator that controls evaluation of its operands.

C. Simple functions

Functions in Velka are created with the special form `lambda`. It creates a function and returns it as a value. The special form `lambda` accepts two arguments: (i) a list of formal arguments and (ii) an arbitrary expression, called body of the function. For example, if we are to create a function which adds two values, we use the following code:

```
(lambda (x y) (+ x y))
```

If a particular type or representation is required. We can specify the type and its representations of the formal arguments of a function. As shows the following example:

```
(lambda ((Int:Native x) (Int:Native y))
  (+ x y))
```

D. Defining types and representations

Definition of a type in Velka is done using the special form `type`. It accepts one argument—a symbol with a name of the type. For example, a set of integers is defined as follows.

```
(type IntSet)
```

Note that type is merely a declaration of a type name. Its semantics is defined further in the program through its representation.

Representations are defined with the special form `representation`. It accepts two arguments—a name of the representation and the name of an existing type. To create representations of a set of integers based on linked lists and bit-vectors one could use.

```
(representation LinkedList IntSet)
(representation BitVector IntSet)
```

Just like the `type` special form, the `representation` special form creates only a declaration. An actual representation of a value is given in a *constructor*, a special kind of function that creates a value and marks it with the representation it is constructing.

To define a constructor in Velka, we use the `constructor` special form. Its first argument is a signature of an atomic representation, the second argument is a list of formal arguments of the constructor, and the last argument is an expression, which returns the constructed value. The value is created with the `construct` special form. For example, the constructor is created and used as follows:[1]

```
(constructor IntSet:BitVector () 0)
(construct IntSet BitVector)
   ;;<0 IntSet:BitVector>
```

E. Deconstruction and Conversions

Naturally, it is necessary to extract a partial information from a value. For this purpose Velka uses the *deconstructors*. The special form `deconstruct` extracts the original value from a value created by a constructor. It accepts two arguments—(i) the deconstructed value and (ii) a representation signature. When evaluated the `deconstruct` special form tries to unpack the deconstructed value into a value with the representation specified as the second argument. If the deconstructed value cannot be unpacked into such value, the execution ends with an error. For example, suppose we want to check if a set, represented by a previously defined `IntSet:BitVector` contains a number 3. We extract the integer value representing the set and then use the bitwise AND operation to check if the third bit is set.

[1]This is a simplified example capable that covers only the values from 0 to 31 for brevity. In practice, a more elaborate solution would be possible and necessary.

```
(define empty-set (construct IntSet BitVector))
(equalp 0
  (bit-and 8
    (deconstruct empty-set Int:Native))) ;;#f
```

Conversions between representations are declared with the special form `conversion`. It accepts four arguments (i) a representation from which a value is converted, (ii) a representation to which a value is converted, (iii) name of an argument, and (iv) body of a conversion. For example:

```
(conversion IntSet:BitVector IntSet:LinkedList
  (IntSet:BitVector set)
  (...))
```

Conversions are performed either implicitly (in most cases) or explicitly by the special form `convert`.

F. let-type and type variables

Velka automatically considers all the symbols in type signatures to be type atoms. To use type variables in type and representation signatures a `let-type` special form must be used. It accepts two arguments—(i) a list of type variable names and (ii) an expression. It allows us to use type variables in the expressions. For example, if we want to make a function that adds an arbitrary value into a collection, we may use the following code:

```
(representation Demonstration List)
(define my-add (let-type (A)
   (lambda ((List:Demonstration l) (A element))
     (tuple element l))))
```

G. Built-in representations and operators

For convenience Velka contains several built-in representations that come from the host environment, in our case from the Java platform. This covers the native representations of primitive data types (integers, floating point numbers, string), lists (ArrayList, LinkedList), tuples, or sets.

Naturally, Velka contains also built-in operators to manipulate these built-in representations. We have already stumbled upon some of these like: `+`, `bit-and`, `tuple`, or `equalp`. It is out of scope of this text to give a full list of these operators. Thus, we explain those used in this text as we go. For the full list of built in operators, please refer to the documentation [7].

III. Extended functions

The key feature of Velka are *extended functions*. The purpose of the extended functions is to bring capabilities of the RDBMS to a general purpose language, allowing Velka to select the most appropriate data representation and algorithm based on the input data. This allows to optimize the program execution without an explicit involvement of a programmer

A. Multiple implementations

Broadly speaking an *extended function* is a function that can select the code to run, depending on the arguments it is applied to. For example, suppose we have an extended function `intersect` that performs an intersection of two sets. We have two representations of sets: the `Set:LinkedList` and the `Set:BitVector` represented by linked lists and bit vectors, respectively.

We can apply `intersect` with either both arguments being a `Set:LinkedList`, both being a `Set:BitVector` or one of each. All of these applications are correct, because the arguments are of the type `Set`. However, they can be of a different representations.

The extended function `intersect` decides during the application which algorithm to use. When both arguments are `Set:BitVector` values, it chooses binary operations. If one argument is a `Set:BitVector` and second is a `Set:LinkedList`, it converts the second argument. If both arguments are `Set:LinkedList`, it uses merge algorithm. Or if the extended function considers it more efficient, it can convert both arguments into the `Set:BitVector`, and use the bit-wise AND operation.

From the technical point of view an *extended functions* can be considered a pair $\langle I, cost \rangle$, where I is a set of simple functions $\{f_1, f_2, \ldots, f_n\}$ called *implementations* and $cost$ is a function that associates each implementation with a cost function. Implementations are algorithms that the extended function chooses from.

All implementations must have the same type; however, they must have a different representation than the other implementations. This ensures that each implementation accepts the same number and types of arguments, and returns a value of the same type. Therefore it can be applied to the arguments of the extended function.

A cost functions $cost(f_i)$ associated with each implementation are responsible for the selection of an implementation applied by extended functions.

B. Running Example

We use type `Name` describing names of persons as a running example in the following text. Depending on the application, in some cases it may be reasonable to represent a given and a family name separately, and in some cases represent them as a single name. Hence, it makes sense to consider a representation `Strd` (structured) consisting of pair of strings, and a representation `Ustrd` (unstructured), consisting of a single string.

In Velka such type and representations are implemented by the following code:

```
(type Name)
(representation Strd Name)
(constructor Name Strd
  ( (String:Native firstname)
    (String:Native surname))
  (tuple firstname surname))
(representation Ustrd Name)
(constructor Name Ustrd
  ((String:Native name))
  name)
```

We assume the following functions for extracting underlying `String:Native` values:

```
(define get-name (lambda ((Name:Ustrd name))
```

```
  (deconstruct name String:Native)))
(define get-given-name
  (lambda ((Name:Strd name))
    (get (deconstruct name
           (String:Native String:Native))
         0)))
(define get-family-name ...) ;; analogously
```

C. Extended functions in Velka

We use the special form `extended-lambda` to create an extended function. It accepts a single argument—a list of type signatures of its arguments. Initially, an extended function does not contain any implementations. For example, consider an extended function `name-equalp` testing equality of names. We start with a definition of an extended function with two `Name` arguments:

```
(define name-equalp
  (extended-lambda (Name Name)))
```

To add a new implementation into an existing extended function, we use the special form `extend`. It accepts three arguments—(i) an extended function, (ii) a simple function, and (iii) a cost function. The cost function, however, is not mandatory. We discuss cost functions in more detail in Section III-F.

For example, if we want to add an implementation for `Ustrd` representation, we use the following code. Also please note, that the `extend` special form does not cause side-effects. It creates a new extended function from the argument, adds a new implementation to it, and returns it. The original extended function is not affected. It is therefore necessary to rebind the symbol `name-equalp` with the special form `define`.

```
(define name-equalp
  (extend name-equalp
    (lambda ((Name:Ustrd first)
             (Name:Ustrd second))
      (equalp (get-name first)
        (get-name second)))))
```

Not any simple function can become an implementation. A simple function must accept the same number and the same type (but not the representations) of arguments as the extended function. Also, the first added implementation sets the return type of the extended function. Therefore any implementation after that must return a value of the same type (but not necessarily of the same representation).

D. Representation of extended function

We cannot use the applicable representation directly to describe a representation of an extended function. Since the implementations can have different representations of arguments and return values.

To resolve these ambiguities we introduce a concept of a *representation set*. It is a finite set of representations, where each representation belongs to the same type. The representation of an extended function is a representation set, containing representation of each implementation.

For example, function `name-equalp` clearly has the type [`Name`,`Name`] → `Name`. However, `name-equalp` contains implementation with the `Name:Strd` and the `Name:Ustrd` representation of the argument. Therefore, its representation is {[`Name:Strd`,`Name:Strd`] → `Bool:Native`, [`Name:Ustrd`,`Name:Ustrd`] → `Bool:Native`}.

E. Application of extended function

Let's say we apply the extended function $\langle I = \{i_1, i_2, \ldots\}, cost\rangle$ on the arguments $a_1, \ldots, a_n$. Each argument a_j is evaluated to the value e_j. Then, the extended function iterates through each implementation i_k. Each implementation i_k has the cost function $cost(i_k)$.

The cost function takes the same type of the arguments as the extended function and returns an integer. Therefore, we can compute cost for each implementation with respect to the arguments, i.e., $(cost(i_k))(a_1, a_2, \ldots)$.

The extended function searches for an implementation i such that, its cost is minimal with respect to the arguments. Then the implementation i is applied as a simple function. For example, let's say we apply our function `name-equalp` with some arguments of `Name:Ustrd` representation.

```
(name-equalp
   (construct Name Ustrd "John Doe")
   (construct Name Ustrd "James Dee"))
```

First both arguments are evaluated, getting values `"John Doe"` and `"James Dee"` both of the `Name:Ustrd` representation.

The exrended function `name-equalp` has two implementations—(i) accepting two arguments of the `Name:Ustrd` representations and (ii) accepting two arguments of the `Name:Strd` representation. We shall call these representations the unstructured implementation and the structured implementation respectively.

Both implementations are defined without a cost function. Therefore, the default cost function is used. The default cost function returns the number of arguments that are not in the expected representation for the implementation.

In our example, the unstructured implementation is expecting two arguments of the `Name:Unstructured` representation. Both arguments have the `Name:Unstructured` representation; therefore default cost function yields 0. On the other hand the structured implementation is expecting two arguments of the `Name:Structured` representation. Since both arguments have the `Name:Unstructured` representation, the cost function yields 2.

The unstructured implementation has the least cost and is applied on the arguments. If there are two or more implementation with the lowest cost, we do not specify which one is selected.

F. Cost functions

A cost function is a function (extended or simple) that is associated with each implementation of an extended function. It computes cost of an implementation for the given arguments. Formally speaking, for the implementation i with the argument

types $t_1, t_2, \ldots, t_n$, a cost function can be any function with the type $(t_1, t_2, \ldots, t_n) \rightarrow$ `Int`.

Let's take the unstructured representation from the previous section. For this implementation a cost function must have the type (`Name Name`) $\rightarrow$ `Int`.

Let's say this implementation is superior in performance, and we want to use it, even if only one of the arguments has the `Name:Ustrd` representation. We use the following cost function:

```
(lambda ((Name:* first) (Name:* second))
  (if (or (instance-of-representation first
             Name:Ustrd)
          (instance-of-representation second
             Name:Ustrd))
    -1 3))
```

This cost function takes two arguments of the `Name` type. Notice that we do not specify the representation of the cost function arguments. Since internally a cost function is applied like any other function, enforcing a specific representation would cause conversion of the representation during application. Therefore special form `instance-of-representation` would not work as expected.

When we apply this cost function with at least one `Name:Ustrd` argument, the function yields -1, which is less than the default cost function yields (the default cost functions minimum is 0). On the other hand, if neither argument is `Name:Ustrd`, the cost function yields 3, which is above what default cost function can yield. Therefore, with this cost function the extended function `name-equalp` selects the unstructured implementation, if at least one of the arguments is `Name:Ustrd`.

To maximize flexibility, we impose minimal restrictions on cost functions. However, this approach introduces challenges as evaluations may fall into non-trivial infinite loops or lead to severe performance loss. User must be aware of these caveats and work accordingly.

IV. Type inference and unification

Velka is a strongly typed functional programming language. The type inference and unification are crucial parts of the language design. Especially, since Velka distincts between types and their representations. In this section we outline the key and the most distinctive parts of the type system. Detailed description is to be discussed in the extended version of the paper.

A. Type and representation unification

Our type unification algorithm is a combination of J. W. Lloyd's [8] and Hindley-Milner's [9] approach. It is explicitly returning either an unification substitution or a `false` answer, while maintaining a predictable and deterministic traversal through a type structure avoiding any side effects during its computation.

The unification of representations is in principle the same as the unification of types. We extended the type unification algorithm for the representation in a straightforward manner. However, we do not introduce representation variables, we use only type variables, which are handled the same way as in type unification. The *representation set* is unique for the representation.

A representation set is meant to represent a set of possible representations of the value. We can unify a representation set with an other representation, if any representation in the set unifies with the other representation.

This trivially holds for a type variable, since it unifies with any representation substituted for it. For an atomic representation, the atomic representation must be present in the set and then they unify with the empty substitution. For functions and tuples, if set contains a representation that unifies over some substitution, the set unifies with the same substitution.

Only case of two representation sets remains. However, we can unify each representation of the first set with the second set. In a recursive call of the algorithm we unify each representation of the second set with the current representation of the first set. Thus we try to unify each representation of the first set with each representation of the second set.

B. Representation inference limitations

Since the representations can be interchanged freely one for another one of the same type (assuming there is a conversion), we cannot unambiguously infer the representation of an expression.

For example, let's consider the `if` expression. In case of the types, the inference rule for this special form is [1]:

$$\frac{c : \{true, false\}, t : A, f : A}{if(c, t, f) : A}$$

Meaning, if we have an expression c of the type $\{true, false\}$ and expressions t and f of the type A, we can infer the expression `(if c t f)` is of the type A.

On the other hand, let's take the Velka expression: `(if (= a 0) 42 (construct Int Roman "XXI"))`. For types everything is clear, the first argument is a boolean and both the second and the third argument are integers. Therefore the expression yields an integer.

However, for representations, the expression yields either an `Int:Native` value or an `Int:Roman` value. At compile time we cannot decide, which integer representation the expression yields.

Therefore, the representation inference rule for the `if` expression is different:

$$\frac{c : \{true, false\}_{Native}, t : A_R, f : A_S}{if(c, t, f) : \{A_R, A_S\}}$$

This way we can unambiguously infer the representation of the expression from the previous example. Since `(= a 0)` has the representation $\{true, false\}_{Native}$, 42 has the representation `Int:Native` and `(construct Int Roman "XXI")` creates a value with the representation `Int:Roman`, the inferred representation of the expression is $\{$`Int:Native`, `Int:Roman`$\}$.

Similar ambiguity is linked with the semantics of extended functions. Therefore, the `extended-lambda` special form and applications of extended functions suffer the same predicament. These are discussed in more detail in Sections V-G and V-H.

V. Representation inference algorithm

An algorithm for representation inference in Velka is designed around the implementation of inference rules for expressions present in Velka. This includes the most frequent language constructs like function applications, tuples, literals, and lambda expressions.

An input of the algorithm consists of two values: e an expression whose representation is inferred and $\mathcal{E}$ an lexical environment, where the expression is inferred.

In this context, an environment is a map of symbols and their bindings. For the inference algorithm, it is important that a symbol is bound to an expression, that infers to the correct representation.

Argument $\mathcal{E}$ influences the inference of symbols as well as functions (since they carry their creation environment). Our algorithm creates a new environment during application inference in a similar manner a lexical closure does.

The inference algorithm returns a pair of values: a representation and a substitution. The substitution carries already inferred or partially inferred type variables over to the further computation. A description of inference rules for Velka expression types follows.

A. Literals and Construct special form

Literal expressions in Velka infers to their assigned representations. Returned substitution is always empty.

The special form `construct` consists of a constructed representation x_r and constructor arguments $a_1, a_2, \ldots, a_n$. The inferred representation is clearly x_r. We must also ensure, that constructor for the arguments exists, and that the arguments are of the correct types.

B. Symbols

When a representation of a symbol is inferred, we inspect if it has a binding in the environment hierarchy. If it has a binding to an expression e_b, we recursively call the inference algorithm on e_b. If a symbol does not have a binding, we cannot infer its representation, therefore we return a type variable and an empty substitution.

C. Substitution Merge

There is a class of type related errors, that occur when a single symbol is used twice, each time as a different type. We need a way to detect this incompatibility between substitutions, and also a convenient way to combine compatible substitutions into a single one. For this purpose we use the substitution merge algorithm.

The merge is based on the idea that two substitutions need to substitute the same variable in order for conflict to arise. If they do not have any common variable, we can use set union to merge them, getting a valid substitution. Note that, even if substitutions have a common variable A, the conflict only occurs if expressions on the right side of the substitution are not equal. Otherwise they are the same expression and set union produces a valid substitution.

Algorithm 1: Substitution merge algorithm

```
1 Function ∪_S(σ, φ)
2     while there is variable A such that A\e ∈ σ, A\f ∈ φ
        and e ≠ f do
3         ρ ← UnifyRepresentations(e, f);
4         if ρ = false then return false;
5         σ ← σρ ;
6         φ ← φρ;
7     return σ ∪ φ
```

Assuming we have two substitutions σ and ϕ, with a common variable A, such that $A\backslash e_\sigma \in \sigma$, $A\backslash e_\phi \in \phi$ and $e_\sigma \neq e_\phi$. If we can find an unifier ρ for e and f, we can compose it with σ and ϕ. In that case $\sigma\rho$ and $\phi\rho$ still have a common variable A, however $\rho(e) = \rho(f)$. Therefore $\sigma\rho \cup \phi\rho$ is a valid substitution.

This is the idea behind the Algorithm 1. We are iterating over common variables that are substituted for not equal expressions of σ and ϕ. For each such variable we find a unifier of the substituted expressions, and compose it with the sets. We end the loop when no such variable can be found and return a set union of the two substitutions. If at any time an unifier does not exists, the substitutions are conflicting and an error is thrown.

D. Special form `if`

We describe the inference of the `if` special form as an example on how a special form representation inference is handled in Velka. Other special forms and tuples are handled in a very similar manner.

The special form `if` takes form of `(if c t f)`, where c is the condition expression, t is a true branch expression and f is a false branch expression. The algorithm infers representations of each of these sub-expressions.

We make sure that the type of c is the boolean and that both t and f infers to the same type. Once the algorithm takes care of this basic type checking, it merges the substitutions of sub-expressions to the substitution for the `if` expression.

The inferred representation is a representation set of representations inferred for t and f, since at compile time it is not possible to decide which one is used. See Algorithm 2 for pseudo-code.

E. Lambda Expressions

A lambda expression assigns representations $r_1, r_2, \ldots, r_n$ to its formal arguments $a_1, a_2, \ldots, a_n$. We use a mock up environment and a special inference-only expressions called *representation holders* to reflect this.

A *representation holder* is a special expression that cannot be evaluated and it infers to an assigned representation with

Algorithm 2: Special form `if` inference algorithm

```
Function InferIfRepresentation(c, t, f, E)
    ⟨r_c, σ_c⟩ ← InferRepresentation(c, E);
    ⟨r_t, σ_t⟩ ← InferRepresentation(t, E);
    ⟨r_f, σ_l⟩ ← InferRepresentation(f, E);
    φ_c ← UnifyTypes(r_c, Bool)
    if φ_c = false then raise error;
    φ ← UnifyTypes(r_t, r_f);
    if φ = false then raise error;
    ψ ← σ_c;
    ψ ← ψ∪_S σ_t;
    if ψ = false then raise error;
    ψ ← ψ∪_S σ_f;
    if ψ = false then raise error;
    return ⟨{r_t, r_f}, ψ⟩
```

Algorithm 3: Lambda expression inference algorithm

```
Function InferLambdaRepresentation([a_1, ..., a_n],
  [r_1, ... r_n], e_b, E)
    Γ ← ⟨{a_i ← e^{r_i} | i = 1, ... n}, E⟩;
    ⟨r_b, σ⟩ ← InferRepresentation(e_b, Γ);
    return ⟨σ([r_1, r_2, ..., r_n]) → r_b, σ⟩
```

the empty substitution. We denote the representation holder's representation using an upper index; for example e^{x_r} or f^A.

When inferring a lambda expression, we create a new mock up environment Γ. Its parent is $\mathcal{E}$, the environment where the lambda expression is evaluated. Γ contains each formal argument a_i bound to the representation holder e^{r_i}. We use the environment Γ to infer the body e_b of the lambda expression. We denote the computed representation r_{body} and the used substitution σ.

It is tempting to use $[r_1, \ldots, r_n] \rightarrow r_{body}$ as the resulting representation. However, consider the following example:

```
(let-type (A) (lambda ((A a)) (+ a 1)))
```

In this case the assigned argument representation is A, r_{body} is `Int:Native`, and σ is $\{A \backslash$`Int:Native`$\}$. Therefore, $[r_1, \ldots, r_n] \rightarrow r_{body}$ is $[A] \rightarrow$ `Int:Native`. However, this lacks already known argument representation.

Thus we use $\sigma([r_1, \ldots, r_n]) \rightarrow r_{body}$ as a resulting representation to propagate information from the body inference. Used substitution is σ. You can see the pseudo code in the Algorithm 3.

F. Unifying representations on a type level

Velka allows to apply a function with correct types, but different than declared representations of arguments, assuming the conversions exists at inference time. Since we infer the representations of the expressions, we need a tool that ensures type safety, and allows different representations.

As discussed in IV-A, to unify types and representations we use the standard algorithm with minor modifications. The algorithm's pseudo-code is presented in Algorithm 4. It accepts representations instead of types and is able to unify representation sets on the type level.

Algorithm 4: Algorithm for type unification working with representations

```
Function UnifyRepresentationsAsTypes(s, t)
    if s and t are representation atoms with the same type
      name or type variables with the same name then
        return {}
    else if s is a type variable then return {s \ t} ;
    else if t is a type variable then return {t \ s} ;
    else if s is a s_1 → s_2 and t is a t_1 → t_2 then
        σ ← UnifyRepresentationsAsTypes(s_1, t_1);
        if σ = false then return false;
        φ ← UnifyRepresentationsAsTypes(σ(s_2), σ(t_2));
        if φ = false then return false;
        return σφ
    else if s is a [s_1, ..., s_n] and t is a [t_1, ..., t_n] then
        θ ← {};
        for i ← 1, ... n do
            ρ ← UnifyRepresentationsAsTypes
                  (θ(s_i), θ(t_i));
            if φ = false then return false;
            θ ← θρ;
        return θ;
    else if s is a {s_1, s_2, ..., s_n} then
        return UnifyRepresentationsAsTypes(s_1, t);
    else if t is a {t_1, t_2, ..., t_n} then
        return UnifyRepresentationsAsTypes(s, t_1);
    else return false ;
```

G. Extended lambda and extend expressions

An extended lambda accepts arguments $a_1, a_2, \ldots, a_n$, with a user assigned types $t_1, t_2, \ldots, t_n$, respectively.

The special form creates an empty container, where implementations are added later. Therefore, we cannot infer a specific representation for the `extended-lambda` special form alone.

We use the representation sets in a form of $\{t_r | t_r$ is a representation of $t\}$, for a type t. This encompasses any possible representation of type t. For convenience we use the following notation: $t^* = \{t_r | t_r$ is a representation of $t\}$.

For an extended lambda `(extended-lambda` t_1 t_2 ... t_n`)` where $t_1, t_2, \ldots, t_n$ are argument types defined by the user, we infer $[t_1^*, t_2^*, \ldots, t_n^*] \rightarrow A$ where A is a new unused type variable. The substitution is empty.

For example, the `(extended-lambda (Int String))` expression infers to the pair $\langle$`[Int* String*]` $\rightarrow A, \emptyset\rangle$.

The other part of the extended functions is the special form `extend`. It has the form `(extend` e_{ext} e_{impl} e_{cost}`)`, where e_{ext} evaluates into an extended function, e_{impl} evaluates into a simple function—the future implementation, and e_{cost} evaluates into the cost function. We put cost function aside for now.

Algorithm 5: Extend special form inference algorithm

```
1  Function InferExtendRepresentation(e_ext, e_impl, e_cost, E)
2      ⟨r_ext, σ_ext⟩ ← InferRepresentation(e_ext, E);
3      ⟨r_impl_L → r_impl_R, σ_impl⟩ ← InferRepresentation(e_impl, E);
4      if UnifyRepresentationsAsTypes
5          (r_ext, r_impl_L → r_impl_R) = false then raise error;
6      ⟨r_cost, σ_cost⟩ ← InferRepresentation(e_cost, E);
7      if unify-representation(r_cost, ⟨r_impl_L → Int:Native)
         = false then raise error;
8      if r_ext has form of [r*_1, r*_2, ..., r*_n] → A then
9          return ⟨{r_impl_L → r_impl_R}, ∅⟩
10     return ⟨e_ext ∪ {r_impl_L → r_impl_R}, ∅⟩
```

Algorithm 6: Inferring result of a function application

```
1  Function ApplicationResultRepresentation
2  ([r_a1, ..., r_an] → r_r, σ_f, [s_a1, ..., s_am], σ_a)
3      ρ ← UnifyRepresentationsAsTypes([r_a1, ..., r_an],
         [s_a1, ..., s_am];
4      if ρ = false then raise error;
5      ρ' ← {A\x | A\x ∈ ρ and A ∉ {r_a1, ..., r_an}} ;
6      φ ← ρ' ∪_S σ_f;
7      if φ = false then raise error;
8      φ' ← φ ∪_S σ_a;
9      if φ' = false then raise error;
10     return ⟨φ(ρ(r_r)), φ⟩
```

Representations r_{ext} and $r_{impl_L} \to r_{impl_R}$[2] are inferred representations of the extended function and the implementation respectively. In the same manner σ_{ext} is a substitution used in the inference of e_{ext}, and σ_{imp} is a substitution used in the inference of e_{impl}.

We check if *types* of r_{ext} and $r_{impl_L} \to r_{impl_R}$ unify. If they do, we add $r_{impl_L} \to r_{impl_R}$ to the set of representations.

A special case arises, if the extended function, does not have any implementation. In that case, the extended function infers to a representation in form $[r_1^*, r_2^*, \ldots, r_n^*] \to A$. This type-wise unifies or not unifies with $r_{impl_L} \to r_{impl_R}$, but we have no set to add the representation to. Therefore, we instead return the singleton $\{r_{impl_L} \to r_{impl_R}\}$.

Since the implementation e_{impl} is specific for certain representation, we cannot use its substitution σ_{impl} for merging. Such merge leads to a conflict, since the arguments of the implementations differ in representations. Therefore, we infer with the empty substitution.

We discuss the cost function now. We make sure that representation r_{cost} unifies with $r_{impl_L} \to$ `Int:Native`. If they do, the cost function has the correct type. If they do not, we return an error. You can see the complete pseudo code in Algorithm 5.

H. Application

The inference of the function application is more complicated than in other languages, due to the presence of extended functions and automatic representation conversions. First, we show an auxiliary algorithm, which is a variation upon the traditional application inference rule [1]:

$$\frac{f : A \to B, x : A}{f(x) : B}$$

Then, we proceed to the main algorithm, which takes Velka's specifics into account.

The auxiliary algorithm (see its pseudo-code in Algorithm 6) is used to infer the representation of the application result, along with the used substitution. It accepts $[r_{a1}, r_{a2}, \ldots, r_{an}] \to r_r$ the applicable representation of the function, $[s_{a1}, s_{a2}, \ldots, s_{am}]$ the representation of the arguments, σ_f the substitution of the function, and σ_a the substitution of the arguments.

We search for a type unifier ρ of $[r_{a1}, r_{a2}, \ldots, r_{an}]$ and $[s_{a1}, s_{a2}, \ldots, s_{an}]$ on line 3, to ensure type safety. We use Algorithm 4, since functions in Velka can be applied with arguments of the correct type and an arbitrary representation. We assume that an arbitrary conversion between representations exists. If ρ does not exists, we return an error.

We cannot use ρ in the further inference. It might introduce an incorrect inference on universally quantified type variables, that are part of the lexical closure. Consider the following example:

```
(define id (let-type (X) (lambda ((X x)) x)))
(tuple (id 42) (id #t))
```

We can easily see that the representation of the function `id` is $X \to X$. In the tuple expression we apply `id` with the `Int:Native` argument, getting $\rho = \{X \backslash$`Int:Native`$\}$. In the second application of `id` we get $\rho = \{X \backslash$`Bool:Native`$\}$ in the same manner.

We omit the rest of the algorithm for now. If we merge the ρ in the substitution of the whole tuple, the two inferred substitutions $X \backslash$`Int:Native` and $X \backslash$`Bool:Native` conflict. But that is not correct, since the type variable X in `id` is universally quantified. Thus such information is excluded from the substitution.

On line 5 we create a substitution ρ' as $\{A \backslash x \in \rho$ such that A is not an universally quantified variable in the $[r_{a1}, r_{a2}, \ldots, r_{an}] \to r_r\}$. This solves the aforementioned issue.

Substitution ϕ aggregates σ_a (the arguments inference), σ_f (the function inference) and ρ' ensuring the substitutions do not conflict.

The inferred representation is $\phi(\rho(r_r))$—an application of the original substitution ρ and the merged substitution ϕ on the right side of the function representation. Used substitution is ϕ.

The application $(f\ \ a_1\ \ a_2 \ldots a_n)$ consists of a function expression f and an argument tuple $[a_1, a_2, \ldots, a_n]$. The inference of the argument tuple yields the representation

[2]We can safely assume, that the representation of e_{imp} have this form, since e_{imp} is a lambda expression by the definition.

Algorithm 7: Function application inference algorithm

Function INFERAPPLICATIONREPRESENTATION(e_f, $[a_1, \ldots, a_n]$), $\mathcal{E}$)
  $\langle r_f, \sigma_f \rangle \leftarrow$ INFERREPRESENTATION(e_f, $\mathcal{E}$);
  $\langle [r_{a_1}, \ldots, r_{a_n}], \sigma_a \rangle \leftarrow$ INFERREPRESENATION ($[a_1, \ldots, a_n]$, $\mathcal{E}$);
  if r_f *is a* $x_r \to y_s$ **then**
    return APPLICATIONRESULTREPRESENATION ($x_r \to y_s$, σ_f, $[r_{a_1}, \ldots, r_{a_n}]$, σ_a)
  else if r_f *is a variable* A **then**
    Let $B \to C$ where B and C are new unused representation variables;
    return APPLICATIONRESULTREPRESENATION ($B \to C$, $\sigma_f \cup \{A \backslash B \to C\}$, $[r_{a_1}, \ldots, r_{a_n}]$, σ_a)
  else if r_f *is in a form* $\{x_{t_1} \to y_{u_1}, x_{t_2} \to y_{u_2}, \ldots, x_{t_o} \to y_{u_o}\}$ **then**
    Let $\{\langle x_{r_i}, \sigma_i \rangle | \langle x_{r_i}, \sigma_i \rangle =$ APPLICATIONRESULTREPRESENTATION ($x_{t_i} \to y_{u_i}$, σ_f, $[r_{a_1}, \ldots, r_{a_n}]$, σ_a)$\}$;
    return $\langle \{x_{r_i} | i = 1, \ldots, o\}, \prod_{i=1,\ldots,o} (\sigma_i) \rangle$
  else raise error;

$[r_{a_1}, r_{a_2}, \ldots, r_{a_n}]$ and the used substitution σ_a. The inference of the function yields the representation r_f and the used substitution σ_f. We discern three cases:

(i) The representation r_f has form $x_r \to y_s$. This is the simplest case. We directly use Algorithm 6 to get a representation and a substitution.

(ii) The representation r_f has form A, where A is a type variable. In this case f is either not bound, or it is a variable of unknown representation. We make new unused type variables B and C, and use $B \to C$ as the argument for algorithm 6. We also add a binding $A \backslash B \to C$ to the substitution σ_f, which is passed to Algorithm 6. This ensures the $A \backslash B \to C$ is passed to the resulting substitution, and the representation, for which A stands, is known.

(iii) In the last case r_f has a form $\{x_{t_1} \to y_{u_1}, x_{t_2} \to y_{u_2}, \ldots, x_{t_o} \to y_{u_o}\}$ of a representation set. We cannot discern which representation is used in the runtime, since the cost function is not evaluated in the inference phase. Therefore, we propagate all possible representations of the result in a representation set.

We call the Algorithm 6 with each representation $x_{t_i} \to y_{u_i}$. We collect the inferred representations to a set and aggregate the used substitutions by the substitution composition. You can see the resulting pseudo-code in the Algorithm 7.

VI. EXPERIMENTAL EVALUATION

We conducted several preliminary experiments in order to measure the performance impact of suggested algorithms and concepts. All experiments were carried out by our Velka implementation [7]. This implementation compiles source code into Clojure [6] source code. This generated Clojure source was then used to run experiments on a computer with two Intel Xeons E5-2680, 64 GB RAM, Debian Linux, OpenJDK 11, and Clojure 1.10.

A. Sorting Implementation

We focused our experiments on an implementation of traditional sorting algorithms. For small data InsertSort algorithm should be faster than QuickSort. Hence, it may be reasonable to switch the sorting algorithm based on the size of input data. In our experiments, we sorted arrays of integers using two representations: `Array:Insertsort` and `Array:Quicksort`. Each representation had its own sorting algorithm using InsertSort and QuickSort, respectively, their detailed description can be found in [10]. The underlying data structure of both representations was a Java ArrayList.

There are three algorithms measured in our experiment. The first is a function *quicksort*, which is a simple function accepting an `Array:Quicksort` argument and uses Quicksort algorithm. The second is *insertsort*, a simple function accepting an `Array:InsertSort` and using InsertSort. The last algorithm—*sort extended* is an extended function accepting any `Array` implementation and using either QuickSort-like divide and conquer recursively calling itself, or InsertSort for small arrays. The divided sub-arrays for QuickSort are eventually sorted using InsertSort once they are small enough. In fact we implicitly obtained a hybrid algorithm.

The threshold for switching from QuickSort to InsertSort was obtained experimentally, by previous experiments with QuickSort and InsertSort algorithms. In the following experiments, the threshold was set to an array of 7 elements, i.e. arrays with 7 or less elements were sorted using InsertSort and larger arrays were sorted using QuickSort. Each algorithm is implemented using an iterative approach in order to get as much performance as possible.

B. Experiments and their results

We sorted arrays of randomly generated positive integers in our experiments. We pre-generated experimental data by a script that uniformly drew numbers ranging from 0 to 9999.

We conducted three batches of experiments to observe the algorithms running on different array sizes. The first batch of experiments is focused on small arrays, up to the 100 elements. The intention is to set the threshold for the *sort extended* algorithm and prove suitability of extended algorithms on a small scale. The second batch of experiments inspects medium-sized arrays ranging from 100 elements to 2900 elements. This experiment intends to compare all three algorithms on a scale, where each one runs in a reasonable time. The last batch of experiments inspects large arrays of integers. It sorted arrays ranging from $200,000$ elements to $500,000$ elements. It intends to compare the performance of large data. All experiments measure time to sort the array in milliseconds.

The results for small data are presented in Fig. 1. Small data were easily handled by each algorithm. Even on the small scale, InsertSort shows worse performance compared to QuickSort and sort extended. The comparison between QuickSort and sort extended is more interesting. You can see the detail of this comparison in Fig. 1 (bottom). We can see, that the two algorithms are very similar in performance.

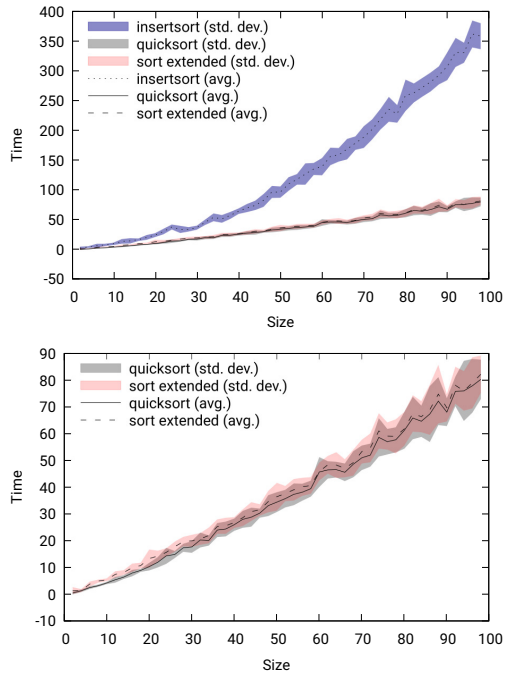


Fig. 1. Algorithm comparison for small data (top), detailed view (bottom)

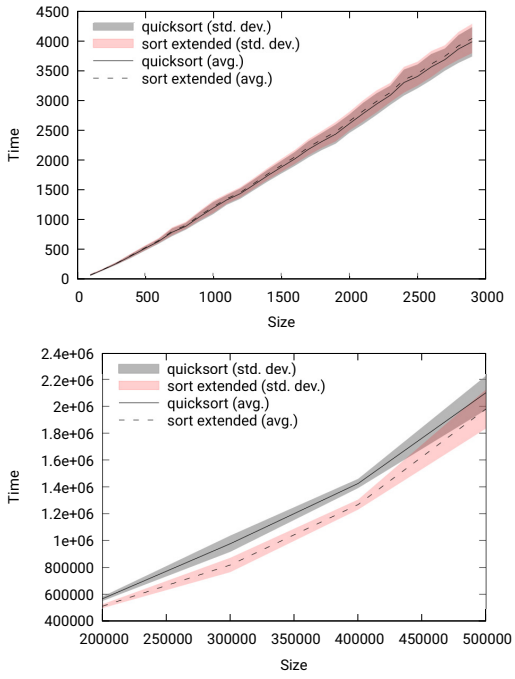


Fig. 2. Algorithm comparison for medium size (top) and large (bottom) data

QuickSort is winning by a small amount. This is probably due to the overhead of an extended function handling, used in sort extended. This suggests that even on a small scale an extended function can be used with only a limited performance loss.

You can see the results for medium-sized data in Fig. 2 (top). The performance of InsertSort is already hindered here. The two other algorithms are almost flat on the x-axis. Therefore, we omit results for InsertSort and present only details for QuickSort and sort extended. The performance gain of QuickSort over sort extended is diminishing. Although QuickSort still performs better, it seems that the overhead of extended functions is reduced due to the performance gain from switching algorithms. The InsertSort is not useful for larger data, and it seems that sort extended starts gaining performance wise on QuickSort and the trend continues.

Results for the experiments with large data are depicted in Fig. 2 (bottom). Comparison between QuickSort and sort extended shows significant development. Sort extended algorithm shows a performance gain around 10% on average compared to QuickSort. It seems, that switching algorithms outweighs overhead, caused by extended functions, on large scales.

VII. Conclusions and Future Research

The Velka language is a framework that allows to dynamically optimize data structures and algorithms to align with input data. We are to extend its standard library to provide various types and representations applicable in data analysis and data management. We are to reduce overhead of the Clojure language by targeting JVM directly. Additionally, we are to explore the possibility of supporting GPGPU computations for matrix and similar types. Further, we aim to research different strategies for setting and fine-tuning cost functions.

References

[1] B. C. Pierce, *Types and programming languages*. MIT Press, 2002. ISBN 978-0-262-16209-8

[2] C. J. Date, *An Introduction to Database Systems, Volume I, 5th Edition*. Addison-Wesley, 1990. ISBN 0-201-52878-9

[3] H. Garcia-Molina, J. D. Ullman, and J. Widom, *Database systems - the complete book*. Pearson Education, 2002. ISBN 978-0-13-098043-4

[4] H. Abelson, R. K. Dybvig *et al.*, "Revised report on the algorithmic language scheme," *High. Order Symb. Comput.*, vol. 11, no. 1, 1998. doi: 10.1023/A:1010051815785

[5] G. Steele, *Common LISP: the language, 2nd Edition*. Digital Pr., 1990. ISBN 0131556649

[6] R. Hickey, "The clojure programming language," in *Proceedings of the 2008 Symposium on Dynamic Languages, DLS 2008, July 8, 2008, Paphos, Cyprus*, J. Brichau, Ed. ACM, 2008. doi: 10.1145/1408681.1408682 p. 1.

[7] R. Skrabal. (2023) Velka source codes. [Online]. Available: https://github.com/Schkrabi/TypeSystem/blob/master

[8] J. W. Lloyd, *Foundations of Logic Programming, 2nd ed.* Springer, 1987. ISBN 3-540-18199-7

[9] R. Hindley, "The principal type-scheme of an object in combinatory logic," *Transactions of the American Mathematical Society*, vol. 146, pp. 29–60, 1969. [Online]. Available: http://www.jstor.org/stable/1995158

[10] D. E. Knuth, *The Art of Computer Programming, Volume III: Sorting and Searching*. Addison-Wesley, 1973. ISBN 0-201-03803-X

Position Papers of the 18th Conference on Computer Science and Intelligence Systems pp. 77–85
DOI: 10.15439/2023F6202
ISSN 2300-5963 ACSIS, Vol. 36

Interconnecting Advanced Networks with AI Applications

Andriy Luntovskyy
0000-0001-7038-6955
BA Dresden Univ. of Coop. Education, Saxon Study Academy,
Hans-Grundig-Str. 25, 01307
Dresden, Germany
Email:
Andriy.Luntovskyy@ba-sachsen.de

***Abstract*— This position/ challenge paper is aimed to explore the potential of advanced network technologies to support Artificial Intelligence (AI) applications and so-called Digital Ecosystems, with a focus on interconnecting both under improving user Quality of Experience (QoE). The work analyzes current opportunities, challenges, and case studies (examples) as well as examines ongoing models and algorithms for future Digital Ecosystems: architectures, platforms, and services. One of the mid-term goals of the author is as follows: to collect further experience together with the fellows and colleagues and to provide the edition of a scientific issue as editor, which is dedicated to the above-mentioned subjects.**

***Index Terms*—Digital Ecosystems, AI, Machine Learning, Neural Networks, 5G and Beyond, Industrial IoT, Software Engineering, Blockchain, CIDN, Honeypotting, Big Data, decision-making, Computer Vision, real-time, Green AI.**

I. Introduction and Challenges

Digital Ecosystems are networked AI services and software platforms for improving user QoS (Quality of Service) and Quality of Experience (QoE). The heterogeneous elements of these systems are interconnected one each other or, correspondently, hierarchically organized and underordered. Each component (software platform, AI, network) in an ecosystem plays a unique role, and they rely on each other for successful operation and survival. This leads to the appropriate ecological metaphor.

Digital Ecosystems, consisting of advanced networking and AI applications, are rapidly changing the world we live in. The influence is bilateral or dual: AI requires the integration of efficient and "green" network technologies and rebuilds the own structure of WWW completely; on the other hand, the sustainable development of network technologies with AI applications is creating new opportunities for enhancing the QoS as well as user QoE in diverse domains, such as every-day workflow and process digitalization, industries, education and high school, ergonomics, sociology, etc. However, these new potentials are also tied to significant challenges related to data mining and security, user privacy, energy efficiency and climate tolerance, and software engineering. This work aims to explore the potentials of advanced network technologies aimed to support AI applications within such Digital Ecosystems, with a focus on interconnecting both under improving user QoE. The work will analyze current opportunities, challenges, and case studies as well as examine ongoing models and algorithms for future Digital Ecosystems: architectures, platforms, and services.

The following subjects are examined below:

- Digital Ecosystems: networked AI services and platforms for improving user QoE
- Next Generation Networks (NGN): 5G and Beyond, Starlink, Terahertz-Band, UWB (Ultra-Wide Band), VLC (Visible Light Communications).
- NGN and AI Applications.
- AI in Digitalization and Industries.
- Industrial Internet of Things (IIoT) and AI.
- AI-based new challenges for Software Engineering: challenges and opportunities for software development with AI.
- Distributed Edge AI: the models on edge devices like base stations, access points, sensors, cameras, IoT devices.
- Applied AI: Machine Learning, Neural Networks, and Deep Learning. Applications and potentials.
- Data Mining and Big Data: Opportunities and challenges for AI in data managing and analyzing.
- AI in Didactics of High School: Benefits and challenges.
- Ethical and Legal Considerations for Generative Language Models.

©2023, PTI

Topical area: Advanced Artificial Intelligence in Applications

- AI for Ergonomics and Sociology: How AI can be used to improve work and living conditions.
- AI and Computer Vision Issues.
- Advanced Security and Ensured User Privacy for AI-based Digital Ecosystems.
- Energy Efficiency and Computational Optimization: Indeed, is AI a climate killer? AI's potential to help reduce energy consumption, improve computational efficiency, minimizing AI's impact on climate change.

II. Advanced Networks Meet AI Platforms

A. Digital Ecosystems

Digital Ecosystems are networked AI services and platforms that improve user QoE. They consist of multiple interconnected devices, applications, and services that work together within modern networks and the internet and provide seamless and personalized experiences to users. Digital Ecosystems are suitable for different domains, such as everyday workflow and process digitalization, industries, education and high school, ergonomics, and sociology. Furthermore, Digital Ecosystems are becoming increasingly important in monitoring and decision-making, justice, marketing, e-commerce, publishing, healthcare, education, arts, and entertainment on an AI basis.

B. Advanced Networking via NGN

So-called Next Generation Networks (NGN) have defined advanced network technologies in the last good 20 years that support higher data rates (DR) and low-latency communication. Some meaningful examples of NGN consider 5G and Beyond, Starlink, Terahertz-Band, UWB, and VLC. These technologies [1-3] are becoming increasingly important, providing Digital Ecosystems with integrated elements of AI as well AI applications themselves. Advanced networking technologies are especially important for AI-supporting applications, which are critical for real-time data processing and rapid decision-making.

C. Applied AI in Digital Ecosystems

Significant role methods and mathematical apparatus such as Machine Learning (ML), Neural Networks (NN), and Deep Learning (DL) based on both above-mentioned approaches are playing (Fig. 1). The combination of them with so-called Language Models gives a demarcation, where we are standing [4, 5]. The typical components of modern AI applications and chat platforms like Chat GPT (Open AI), Bing (Microsoft), Bard AI (Google), Meta's Platforms Chatbot (without FB and Instagram), Chinchilla (DeepMind), and, furtherly, Jasper, Quillbot, Bloom, Replika, ELSA, Bing AI, Dall-E are as follows [4-11]:

- Transformers: originally from Google Brain, trained with so-called Reinforcement Learning.
- RLHF (Reinforcement Learning from Human Feedback).
- PPO (Proximal Policy Optimization).
- GUI/ user interface for text input and output.
- Language models (refer to Fig. 2), such as Model OpenAI's GPT-3.5/ GPT-4, LaMDA Google Language Model, and LLaMa (Large Language Model of Meta AI).

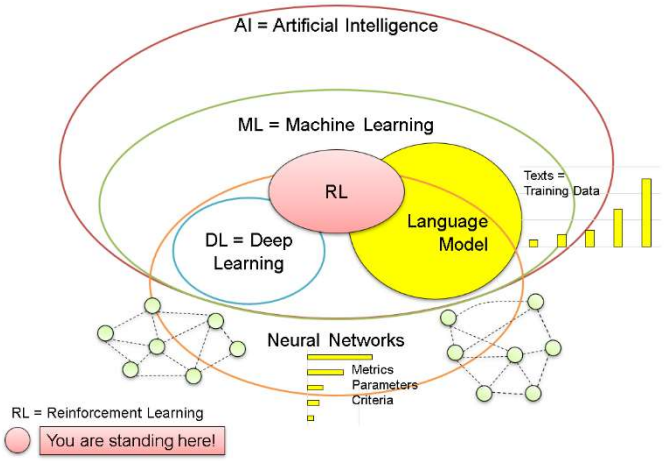


Fig 1. Demarcation of AI methods in Digital Ecosystems

Practically, model training data for the language models consists of a large number of texts created by humans, communities, and some non-profit institutions, e.g., commoncrawl.org as a text's source can also be specified as the so-called C4 (Common Crawl Cultural Context). Furthermore, the sources of texts can be differentiated as follows: search engines, online forums and communities, social media, blogs, newspapers, scientific articles, online books, wikis and encyclopedias, GitHub, spoken language, videos, and audio materials, etc. (refer to Table I).

TABLE I.
DATA SETS FOR AI TRAINING [11]

Data Set Examples	Usage in %
Common Crawl	67
C4	15
GitHub	4.5
Wikipedia	4.5
ArXiv	2.5
Stack Exchange	2.0
Other	4.5
Overall:	100

The well-known applied AI (refer to Table II) is dedicated to modern healthcare process management, finance and credit policies, logistics, and transportation. Some examples of Applied AI include medical diagnosis, report completion, analysis and prohibition for security denials, collections of laws in jurisprudence, and fraud detection. The used language models can be trained with huge amounts of different parameters:

from 7 up to approx. 70 billion of parameters, such as for GPT-3.5/ GPT-4, or LLaMa (refer Fig. 2).

D. AI in Digitalization and Industries

AI platforms are becoming of great importance in modern digitalization processes and in industries. AI and, especially, ML (Machine Learning) applications (refer to Fig. 3) support process automation and improve human decision-making based on often repetitive routines. Some examples of AI in digitalization and industries include chatbots, predictive maintenance, and autonomous vehicles (UAV – unmanned automotive or aerial vehicles). ML can be differentiated into three known types: Supervised Learning (SL), Unsupervised Learning (UL), and Reinforcement Learning (RL), as it was shown in Fig. 3, where RL is the most used type for the above-mentioned language models [4-7].

A special part, Semi-Supervised Learning (SL/2), is an essential ML method whose importance has increased with the deployment of LLMs in recent years. Intuitively, SL/2 can be viewed as an exam, data can be viewed as problem examples that the teacher solves for the class to help solve a different set of problems. The unresolved problems act as further exam questions or they become the practical tasks that make up the exam. It looks like data clustering and then labeling the clusters with labeled data, moving to the decision boundaries. SL/2 is without scope in this taxonomy, but important for LLMs.

E. Industrial Internet of Things and AI

The Industrial Internet of Things (IIoT) refers to the use of connected devices like base stations, access points, sensors, cameras, and IoT devices in industrial environments (Digital Ecosystems). Data acquisition and analysis are provided in the industrial scenarios, overcoming and avoiding "Big Data" bottlenecks for IoT devices with valuable insights and improving operational efficiency [1-3].

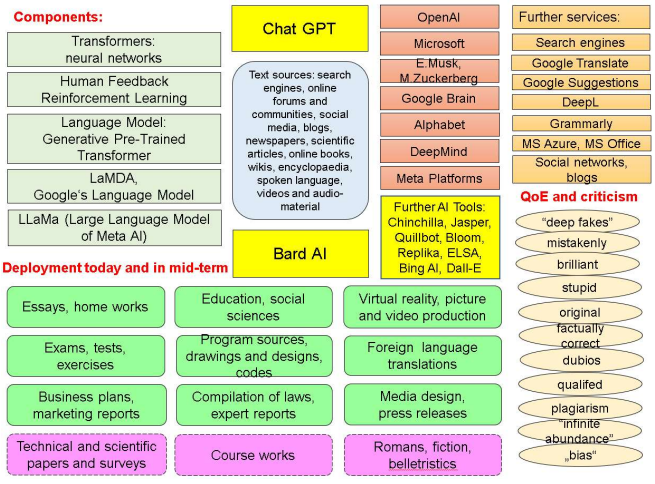


Fig 2. AI platforms nowadays

F. AI-based Software Engineering

The integration of AI in software development presents new challenges and opportunities. On the other hand, there many risks are tied to it. These challenges include necessarily human-AI collaboration in software engineering practices as well as require appropriate tools which boost the control on possible bias and "human factors", e.g., advanced GIT or GitHub.

Such integration of AI and GIT/GitHub can help to improve the quality and efficiency of software development, high re-usability grade of mistake-free source code fragments, embedding to so-called agile process models [3] under the use of ML-like XP, Scrum, consecutive providing of some software engineering techniques like, e.g., agents or micro-services. By AI-supported version and data controlling, developers can better track changes over time and collaborate more effectively with other developers (DevOps).

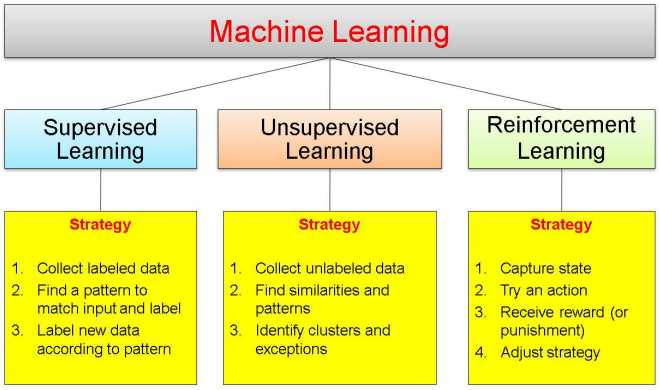


Fig 3. Classification of Machine Learning [4,5]

G. Ethical and Legal Considerations for Generative Language Models

The coexistence of AI and networks in modern Digital Ecosystems produces the need for deepening ethical and legal considerations ("robots may take jobs from humans away"). Also, there are some anxiousness and troubles concerning AI's role in curricula and didactics of school education and high school. Some evident benefits and potentials in the rapid development of cognitive skills can be accompanied by big risks and challenges ("digital dementia"). AI can also be used in education to improve the teaching and learning experience. AI can help to personalize learning, provide feedback, and enhance student engagement. So, there are also ethical, legal, and privacy limitations as well as further well-known risks for AI platforms. The continued development of AI technologies has already led to the possibility of automating some repetitive and routine tasks, including potential replacement in the mid-term for the professions mentioned in the following list: foreign language translator, exercise tutor, data input specialist, helpdesk consultant, proofreader, lawyer assistant, ledger accountant, advertisement and marketing specialist, and copywriter [4-17]. Even more, AI will replace long-term perspective market research analysts, social media managers, meeting planners, telemarketers, virtual assistants, audio-2-text typists, journalists-reporters, travel agents, tech support analysts, content moderators, and personnel recruiters. However, it's important to notice that the capabilities of AI platforms

like Chat GPT, Bard AI cannot yet provide advanced functionality aimed at fully replacing complex human tasks under existing legal considerations (legislation and laws) anytime soon. In addition, according to the experts, critical professions that mostly require the highest level of accuracy and on which human life depends will anyway survive [4-17].

III. Well-known Risks for AI Platforms

A. Awareness of the Significant AI Risks

This is one of the most important topics for Digital Ecosystems. The Top10 of these risks are as follows [12]:

1. Plagiarism culture
2. Copyright issues and violence
3. Free user contributions with middle to low quality
4. Information but not knowledge
5. Intellectual stagnation
6. Data security and user privacy
7. Unrecognized bias (systematic error)
8. Human credulity
9. Digital dementia
10. Mythos about "Infinite Abundance".

Such AI risks can be considered nowadays on the example of the most known platform Chat GPT (refer Fig. 4) as well as on the different AI systems like Bard AI, Chinchilla, Meta's Platform Chatbot, etc. (refer Table II).

Example 1:

OpenAI was founded in 2015 as a non-profit research and development organization by Tesla and PayPal (inter alia CEO Elon Musk, CTO Sam Altman). The AI platform from OpenAI is titled Chat GPT and possesses the following architecture (refer to Fig. 4) and distinguishing features [4-10]:

- A free license with a limited knowledge base by the year 2021 is available
- Furthermore, a paid monthly subscription for Chat GPT Plus and a professional version
- Further integration follows in multiple MS products in the mid-term such as Word, Excel, Bing, Edge, and Azure (refer to the given ecosystem)
- Availability as a standalone or web app, however, has a lot of risks due to fake accounts and malware attempts!
- But that's not all: Chat GPT can impress with other parameters. According to the estimations, the AI knows the data sets for approx. 570 gigabytes (refer to Table I and Table II). The vocabulary is spread over 95 different languages [8-11]; however, access to the software can be restricted for selected countries
- Additionally, the chatbot provides multiple programming languages, including Python, JavaScript, C++, and SQL.

B. An Effect titled "Bias Forever"

Bias or distortion as a systematic error of an estimator is in the estimation theory. That distortion and bias are a problem in AI models is nothing new. These phenomena are known as GIGO ("garbage in, garbage out"): if the acquired datasets are inaccurate or "biased", this will be reflected in the results. Even hardware and logic output (AI engines) can work correctly and logically, but humans cannot. This is especially dangerous due to the ever-closer integration of AI techniques into existing search engines, clouds, office software, IDEs and compilers, and groupware (e.g., such as from Microsoft and Google).

C. Distributed Edge AI and Big Data

Distributed Edge AI refers to the deployment of AI models and methods (like the above-mentioned ML, NN, DL) on networking edge devices like base stations, access points, sensors, cameras, and further IoT devices. Distributed Edge AI can help to reduce latency and improve the efficiency of decision-making in real-time in Digital Ecosystems.

Data Mining and Big Data refer to the collection, processing, and analysis of large amounts of data. AI can help to examine and build primary clusters of "Big Data", providing valuable insights and improving decision-making [1-3].

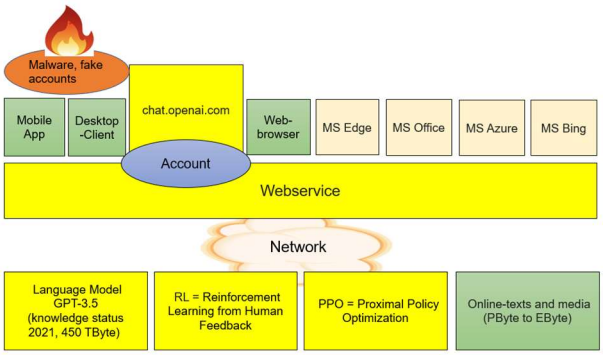


Fig 4. Chat GPT at a glance and its ecosystem

IV. AI and Computer Vision Issues

A. How does Computer Vision work?

AI, per definition, is a complex of multiple models and methods to support computers to become able to solve logical problems on their own. ML "learns" using acquired examples or training data. We speak about Deep Learning (DL) when the learning process is only made under the support of neural networks (NN).

Machine Vision deals with image acquisition and processing to generate added value. So-called Computer Vision (CoVis) is a part of AI and refers to the methods used by DL and Machine Vision (refer to Fig. 5).

Hence, in the opinion of some researchers, Computer Vision as a discipline contains many further methods and algorithms that are not part of Deep Learning: filtering, Fourier and wavelet analysis, etc. Furthermore, Computer Vision is an application of the DL for solving Machine Vision tasks. It's a three-part process (refer to Fig. 6):

1. A model is "trained" on the basis of available training data.
2. The model performs later a Computer Vision task.
3. The graphical objects are recognized with a high probability.

B. Metrics and Confusion Matrix

So-called metrics are the model variables and numerical values used to measure the quality of the deployed models for ML, DL, Machine Vision, and Computer Vision [1-5]. A certain metric can be identified under the use of the so-called Confusion Matrix (Table III). Such a matrix can record how many errors and hits the discussed model has when the model is executed.

The usage purpose for so-called metrics is as follows:

- Understandable criteria for comparing processes and products within a discussed model can be provided.
- Quality and model properties can be quantitatively described via the metrics.

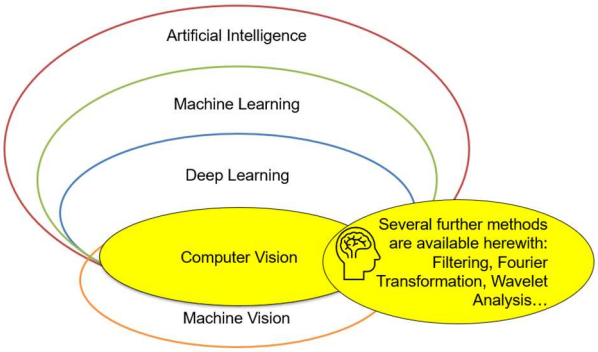


Fig 5. Demarcation CoVis within AI

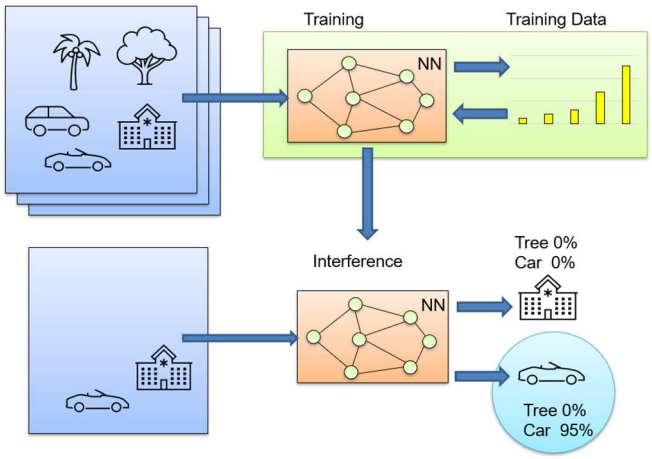


Fig 6. How does Computer Vision work?

TABLE II.
AI PLATFORMS IN COMPARISON (UNDER USE [6, 7], STATUS ON APRIL 2023)

Features	AI platforms		
	Chat GPT	***Bard AI***	***Chinchilla***
Release dates	November 2022	February 2023	March 2022
Vendor	Open AI	Google/ Alphabet	DeepMind (Google's AI firm)
Status	Free version and subscriptions	Limited availability (currently only available to testers)	Presently unavailable to the public
Construction	Open-source conversational AI chatbot Built on Open AI's GPT-3.5/ GPT-4 Draws information from data it's trained on	AI-powered conversational chatbot Built on Google's LaMDA Draws information from the internet	Made with 70 billion parameters
Components and particularities	Transformers with an original ancestry from Google Brain, trained with so-called RLHF (Reinforcement Learning from Human Feedback) PPO (Proximal Policy Optimization) Text input and output with language model GPT-3.5/ GPT-4	Transformers are based on a networked neural model LaMDA, Google-language model Google Research elaborated these transformers even in 2017 Surprisingly, the transformers and the GPT-3 language model use both the basics for Chat GPT too!	Transformers of DeepMind are used
Deployment	Chatbot functionality Generating reports and summaries Creation of marketing material Performing language translation Providing ideas Elaboration of source code and essays Explanation of complex concepts Providing virtual assistants	Assistance in basic search functions Explanation of complex concepts Providing ideas	Chatbot functionality Providing virtual assistants Use of predictive models Creating video game characters Improvement of digital products
Performance and QoE	middle	good	middle
Chatting language support	95 natural languages, among them English, Spanish, French, German, Japanese, Ukrainian	Only English	Only English
Support of programming languages for source code generation	provides multiple programming languages, including Python, JavaScript, C++, and SQL	multiple programming languages	multiple programming languages

Example 2:

There are n=165 elements in an investigated party; they are distributed according to a binary criterion (e.g., for red= yes, for blue= no). Based on a given CoVis method, we have obtained 55 predicted blue and 110 predicted red elements. Indeed, there is the following distribution: 60 blue and 105 red real elements in the party in fact (refer to Table III).

Therefore, we can differentiate here TP (true positive), TN (true negative), FP (false positive), FN (false negative) elements for a given CoVis method.

TABLE III.
CONFUSION MATRIX

n=165	*Predicted NO*	*Predicted YES*	
Fact NO	TN=50	FP=10	60
Fact YES	FN=5	TP=100	105
	55	110	

V. ADVANCED SECURITY AND ENSURED USER PRIVACY FOR AI-BASED DIGITAL ECOSYSTEMS

The integration of AI in Digital Ecosystems provides advanced networking security and user privacy challenges.

The functionality of FW is growing each decade, including IPS, IDS, Antibot, CIDN, and other concepts. Blockchain technology substitutes well-known PKI infrastructure and meets AI (refer to Fig. 7).

AI supports cybersecurity frameworks using OWASP, SIEM, and MITRE foundations [29-32] as training data sets.

A. Advanced Firewall Techniques

There are multiple opportunities as follows [1-3]:

- Advanced firewall techniques must be used too, like IDS/ IPS (Intrusion Detection and Prevention), and CIDN (Collaborative Intrusion Detection Networks), to secure against more and more sophisticated intruders and insider attacks.
- Firewall techniques and advanced CIDN can be boosted via available AI methods.
- Honeypots provide the decoys for detracting multiple intruders and insiders from real attack targets. The diversity of honeypots with collecting knowledge about dangerous events plays a steadily growing role in secure networking.
- Honeypotting role can be even more increased by the deployment of a trendy "deception technology" with AI elements.

B. Blockchain Issues for AI

Blockchain (BC) and Smart Contracting (SC) based on BC provide the compulsoriness in decentralized communication scenarios like up-to-date Peer-2-Peer and Machine-2-Machine must be deployed instead of convenient PKI (Public Key Infrastructure) infrastructures for the discussed AI-based digital ecosystems [1-3].

They need additional so-called NFT (non-fungible tokens), which provide a public certificate of authenticity, authorship, or proof of ownership of a digital media file (digital artifact or asset). What does it mean for AI and Digital Ecosystems?

NFTs are created under the use of existing blockchains and combining records containing cryptographic hashes which uniquely identify a set of data. An NFT is a cryptographically secured unit of data stored on the BC that can be sold and purchased on the digital markets, analogically to crypto-currency such as BTC or ETC.

An NFT may be associated with a specific digital asset, such as an image, art piece, music work, sports events, or an AI-generated product, and may grant licensing rights to use such digital media file (digital artifact or asset) for a specific purpose.

An NFT does not restrict the sharing or copying of the associated digital files, nor does it prevent the creation of a new NFT. An NFT usually results in an informal exchange of ownership of an asset.

Often, an NFT doesn't possess absolute legal rights; they are frequently absent in the actual legislation. Therefore, NFTs' use is uncertain under applicable law and frequently has an extra-legal character. Hence, ownership of an NFT often has no legal meaning and does not necessarily protect the copyrights and intellectual property rights in the associated digital files.

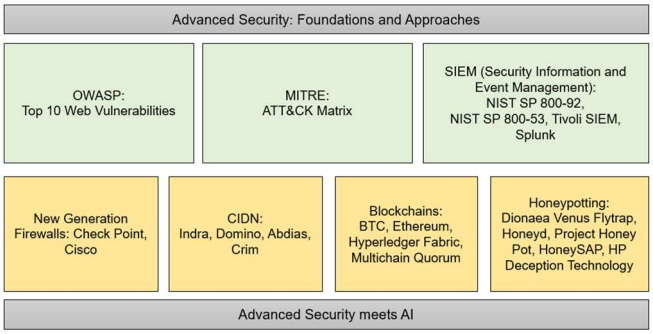


Fig 7. AI-boosted Advanced Security:
Foundations and Best Practices [1-3, 29-32]

Example 3:

NFTSs act herewith like digital certificates of authenticity. They ensure that among a large number of entirely identical copies of a media file or a digital artifact, only one file can be considered a signed, compulsory original. The NFT hype peaked in 2021 for human-made digital artworks and continues for AI-generated digital paintings nowadays.

There are a lot of AI-based Text-to-Image-Generators nowadays:

- Open-source Text-2-Image-Gen: Stability AI (Stable Diffusion)
- Commercial AI tool: Dall-E (OpenAI).

An AI software called "Female Artist Claire Silver" first announced in March 2023 its new collection, "Can I tell you a secret?", which obtained a premiere at the Louvre in Paris. One of the titles is: "Love in the 4th Turning". Overall, the collection consists of 100 paintings secured via Superchief NFT Gallery in NY. On Twitter: Claire Silver would like to "…bring AI art to mainstream culture together".

VI. AI IN DIDACTICS OF EDUCATIONAL INSTITUTIONS

The challenges and risks of this topic are as follows [13-17].

1. Teaching media skills: One of the most important tasks of educational institutions (schools, high schools, universities) should be to train learners in the use of digital media and AI. They should learn the critical answering approaches, whether a text was generated by a human or by an AI.

2. Integrating AI technology into teaching: AI will play an increasingly growing role in the mid-term, so educational institutions should incorporate AI technology into the curriculum to prepare the pupils and students for the demands of a rapidly changing labor world. For example, chatbots or language assistants can be used in course works and class exercises.
3. Generation of teaching materials: Generally, AI text generators can be a valuable addition to teaching materials if used appropriately. However, universities should ensure that the generated materials are of appropriate quality, support the curricula, and are regularly reviewed to ensure timeliness and relevance. In addition, ethical and legal aspects for educational institutions should be considered, and the generated materials should always be used only as a complementary (secondary) to the materials produced by teaching staff.
4. Training and guidelines for teaching staff, pupils, and students are required, which are aimed at avoiding plagiarism under the use of AI generators. The test and exam regulations should include clear requirements for training and policies that ensure all stakeholders understand the risks and consequences of plagiarism and how to identify AI-generated theses, course works, and essays.

Example 4:

Some useful tools can be cited herewith, like GitMind, Miro Mind, or OrgPad [18-21], which are deployed to acquire and shape diverse ideas, hypertext links, networked media, and further essential entities to existing practical areas (domains). Such preliminary AI tools support so-called ontologies, mind maps, namespaces, and source code fragments in a target metalanguage (OWL, XML, JSON, etc.) to certain domains, like medicine, software engineering, or (high-school) didactics, but are augmented via the linked multimedia sources, which can be collected via internet, on online communities, social media, blogs, newspapers, scientific articles, online books, wikis, etc.

GitMind [18] is an online mind mapping and flowchart tool. With GitMind, users can easily create diverse mind maps, flowcharts, and UML diagrams, providing useful domain modeling with AI elements.

Miro Mind Mapping Tool [19] provides an elaboration of different mind maps or ontologies under comfortable GUI. This leads to accelerate source code development, boosting, and embedding of mistake-free source code fragments.

OrgPad [20, 21] is directly dedicated to mind maps and is simultaneously a flowchart tool with the support of hypertext links and a comfortable as well as understandable GUI. Such kind of tools can be programmed under the use of LISP or modern LISP dialects like Clojure or ClojureScript, based on JavaScript. They are frequently used as a first stage for entering, shaping, and saving the facts and knowledge into KDB (Knowledge Data Bases), which are used in AI applications (for Logical Output, NN, Fuzzy Logic, Big Data, Decision Making Support).

VII. AI Tools "at Glance": Advantages and Criticism

A. Energy Efficiency and Computational Optimization

Indeed, is AI a climate killer? Is it possible, a "Green AI"? Such statements are rather incorrect, but the intensive use of AI applications can surely increase the CO_2 impact and commonly contribute to the greenhouse effect [1-3]. AI's potential can rather be used to help reduce energy consumption, improve computational efficiency, minimizing of AI's impact on climate change.

B. Some Advantages for Insurers and Recruiters

Let's consider the following examples:

Example 5:

Insurance companies are steadily investing nowadays in new services, chatbots, and online claims processing [22]. In the mid-term, there are significant changes for the customers too. Through AI and generative language models, customer communication becomes more flexible and more personalized, from the signing of the polis to the processing of led claims. Like at the airport or in the supermarket, where employees support customers digitalizing their self-check-in or scanning the barcodes for goods, insurance companies' collaborators provide simple products to complete them online and show how easy it is, standing out of scope but implicitly involved.

Example 6:

Didn't Chat GPT formulate your CV and advertisement letter? This is how an HR collaborator (recruiter) normally reacts to an AI-manipulated advertisement letter. A lot of Chat GPT formulations are still dubious or clumsy in the meantime but provide a good perspective. In some places can be remarked that there is too much direct translation from English into German or else, from analogous languages. Nevertheless, you can use a Chat GPT-generated text for an advanced electronic final advertisement letter enclosed to your CV and the digitalized scans of your certificates. And a good result can be indeed shown. The generated texts are sometimes surprisingly good, even better than the recruiters can expect from a human candidate who evidently knows the workflow perfectly [23].

C. Warnings Against Generative AI-Language Models

An updated survey from the USA confirmed that 43% of all employees have already used Chat GPT at the workplace. Unfortunately, as this new survey has shown, too, the AI tools provide not only positive effects. Even in contrast: it could affect up to 300 million jobs worldwide, especially lawyer assistants and management employees [24].

An open letter published by the non-profit society "Future of Life Institute", has already been signed by approx. 34,000 prominent people, including Elon Musk and Steve Wozniak [25-27]. E. Musk: "Powerful AI systems should not be developed until we are confident that their impact will be positive and their risks manageable". The open letter also mentions the potential risks to our civilization from generative AI language models in the form of economic and political disruption (fake

news, manipulated societies and finances, uncontrolled access to dangerous substances, weapons, and cyberwars).

The letter was signed also by British astrophysicist Stephen Hawking among other prominent persons. “Success in the creation of artificial intelligence may be the greatest achievement in the history of human civilization. But they may be the last if we do not learn to avoid risks, ” - said Hawking. This letter was also supported by several human rights organizations, such as Human Rights Watch.

The EUROPOL had already warned of the possible misuse by phishing attempts, disinformation, and cybercrimes. Herewith some conclusions from the recent EUROPOL reports are cited: AI makes it easier for criminals to misuse available knowledge for malicious purposes such as illegal weapon operations, terror, porn and sex, drug-pushing, networking, and software hacking, but it also provides for law enforcement officers, agencies and forensics new ways to fight against such crime challenges [24-27]. E. Musk: "AI is a potentially existential threat to humanity. A training pause for AI is required" [25]. Autocratic regimes can use AI for the production of fake news and for propaganda purposes. Some such states regard AI as strategically very important and would like to give AI researchers multiple freedom grades. For this aim, domestic access to AI tools must be critically restricted (s. Wassenaar Arrangement - 1996) [28].

VIII. Increasing AI Networking Efficiency

LLMs are real resource consumers because of the huge computing power that the highly-performant parallel clusters must perform in the background and, therefore, require huge amounts of electrical energy. To enhance the efficiency of existing AI infrastructure, leading high-tech companies are deploying their AI-optimized approaches, including the implementation of liquid-cooled hardware and high-performance AI networks. These networks interconnect ~10.000 specialized AI chips within large-scale training clusters, enabling faster data processing and reducing costs associated with data center construction [1-3]. The efficiency of such clusters can be measured in terms of PUE and ERE [1-3, 47-49]:

$$1 < PUE = (P_A + P_{IT})/ P_{IT}$$

$$1 < ERE = (P_A + P_{IT} - P_R)/ P_{IT}$$

where PUE – Power Usage Effectiveness, ERE – Energy Recycling Efficiency, P_A – is waste power, P_{IT} – is raw IT service power, P_R – is recycling power. The best practices have shown:

$$1 < PUE < 1.16 \qquad 1 < ERE < 1.06$$

Examples of such efficient clusters are multiple [47-49].
Aimed at enhancing the efficiency of existing AI infrastructure, leading high-tech companies are deploying their AI-optimized approaches, including the implementation of liquid-cooled hardware and high-performance AI networks.

These networks interconnect ~10.000 specialized AI chips within large-scale training clusters consisting of multiple boards, enabling faster data processing and reducing costs associated with data center construction. The AI networking innovations can be divided into three strata (Fig. 8):

- Cluster stratum.
- Board stratum.
- Processor unit stratum.

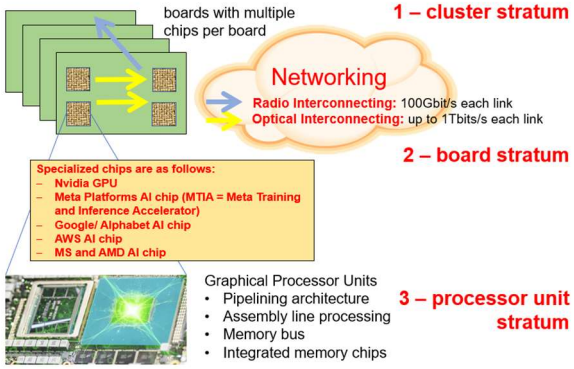


Fig 8. AI networking and computing innovations

The cluster stratum is represented via liquid-cooled, high-performance, and large-scale training clusters [33-35]. On the board stratum, the performant optical and radio links are deployed on the short range between the chips and boards. High-efficient precision antenna constructions (like 3D-MIMO) as well as Terahertz-solutions are used [1-3].

IX. Conclusions

1) This is a work-in-progress. Based on the discussed aspects a new Springer LNEE series title is prepared for the year 2024: “Digital Ecosystems: Interconnecting Advanced Networks with AI Applications” with approx. 700 pages, 39 chapters each up to 16-20 pages in Springer format, under our own edition, in close cooperation with scientists from Germany, Ukraine, Switzerland, Slovakia, Czechia, Poland, China, Italy, Azerbaijan, France, Cameroon, United Kingdom, and many others. This book will continue the former Springer LNEE book series [1, 2]:

- Andriy Luntovskyy, Mikhailo Klymash, et al. (Eds.). "Intent-Based Networking" (2022, LNEE 831, ISBN 978-3-030-92433-1).
- Andriy Luntovskyy, Mikhailo Klymash, et al. (Eds.). "Emerging Networking" (2023, LNEE 965, ISBN: 978-3-031-24962-4).

2) Digital Ecosystems are networked AI services and platforms that improve user QoE. They consist of multiple interconnected devices, applications, and services that work together within modern networks (5G and Beyond, optical networks, WLAN, Industrial IoT, UWB, Starlink).

3) Digital Ecosystems are becoming increasingly important in monitoring and decision-making, justice, marketing, e-commerce, publishing, healthcare, education, arts, and entertainment on an AI basis. The development of so-called "AI co-pilots" for office apps, teamworking,

IDE/GIT, as well as malware defense tools will significantly reduce the risks in the short term and increase the overall usability of generative language models.

4) Distributed Edge AI in Digital Ecosystems refers to the deployment of AI models and methods (like the above-mentioned ML, NN, DL) on networking edge devices like base stations, access points, sensors, cameras, and further IoT devices. Distributed Edge AI can help to reduce latency and improve the efficiency of decision-making in real-time in Digital Ecosystems.

5) New challenges in Software Engineering include human-AI collaboration, require appropriate tools which accelerate software development based on agile process models, and boost the control of possible bias and "human factors".

6) The integration of AI in Digital Ecosystems provides advanced security and privacy challenges. There are multiple opportunities like Blockchain (SC, NFT), CIDN, and Honeypotting in decentralized communication scenarios like up-to-date Peer-2-Peer and Machine-2-Machine.

7) Educational institutions should integrate AI technology into the curricula to prepare pupils and students for the demands of a rapidly changing labor world. Training and guidelines for teaching staff, pupils, and students are required, which are aimed at avoiding plagiarism under the use of AI generators.

8) This is a position/ challenge paper. One of the mid-term goals of the author is as follows: to collect further experience together with the fellows and colleagues and to provide the edition of a scientific issue as editor, which is dedicated to the above-mentioned subjects.

ACKNOWLEDGMENT

The author's acknowledgment belongs to the colleagues and fellows from Dresden, Lviv, and Prague: Steffen Greiffenberg, Frank Schweitzer, Dietbert Guetter, Juergen Smettan, Tenshi Hara, Mykola Beshley, Mykhailo Klymash, Bogdan Shubyn, and Adam Kalisz for valuable impulses by fulfilling this work.

REFERENCES

[1] Andriy Luntovskyy, Mykola Beshley, Mikhailo Klymash (Eds.). Future Intent-Based Networking: on the QoS Robust and Energy Efficient Heterogeneous Software-Defined Networks, in LNEE 831, 28 chapters, monograph, on 15.01.2022, Book, Hardcover, XXI, 530 pages, 1st ed. 2022, Springer International Publishing (ISBN: 978-3-030-92433-1).

[2] Andriy Luntovskyy, Mykola Beshley, Igor Melnyk, Mykhaylo Klymash, and Alexander Schill (Eds). Emerging Networking in the Digital Transformation Age: Approaches, Protocols, Platforms, Best Practices, and Energy Efficiency, by Springer LNEE 2023, issue 965, Springer Nature Cham, Eds. Andriy Luntovskyy, Mykola Beshley, Igor Melnyk, Mykhaylo Klymash, and Alexander Schill (ISBN: 978-3-031-24962-4), 2023, XXXVI + 660 pages, 335 illus., 208 illus. in color.

[3] Andriy Luntovskyy, Dietbert Guetter. Highly-Distributed Systems: IoT, Robotics, Mobile Apps, Energy Efficiency, Security, Springer Nature Switzerland, Cham, monograph, ISBN: 978-3-030-92828-5, 1st ed. 2022, XXXII, 321 pages, 189 color figures (Foreword: A.Schill).

[4] S. Seegerer, T. Michaeli, R.Romeike. So lernen Maschinen, 2020, pp. 27-31.

[5] T. M. Mitchell. Machine Learning, Boston, McGraw-Hill, 1997.

[6] PC Guide: Bard AI vs. Chat GPT (online): https://www.pcguide.com/apps/bard-ai-vs-chat-gpt/.

[7] PC Guide: Chat GPT vs. Chinchilla AI (online): https://www.pcguide.com/apps/chat-gpt-vs-chinchilla-ai/.

[8] R. van Root. E-Books von Chat GPT tauchen bei Amazon auf – Problem für Verlage (online): https://OpenAIunsplash.com/.

[9] F. Peters. E-Books von Chat GPT tauchen bei Amazon auf – Urheberrecht unklar. Basic Thinking (online): https://www.basicthinking.de/.

[10] Sam Altman. Chat GPT: Open AI-Gründer Sam Altman fordert Regulierung von KI (online): https://OpenAIunsplash.com/.

[11] Metas KI Modell LLaMA wurde schon geleakt (online): https://www.computerwoche.de/a/metas-ki-modell-llama-wurde-schon-geleakt,3614004?xing_share=news.

[12] P.Wayner. Angst x Chat GPT & Co.: Zehn Gründe, Generative AI zu fürchten (online): https://www.computerwoche.de/ap/peter-wayner,3298

[13] Einordnung von Chat GPT, Chancen und Risiken u. Einsatzmöglichkeiten in der Lehre, Universität Hamburg (online): https://www.hul.uni-hamburg.de/selbstlernmaterialien/dokumente/hul-chatgpt-im-kontext-lehre-2023-01-20.pdf.

[14] Gestaltung der Hochschullehre mit KI, TU Berlin: https://www.tu.berlin/bzhl/ressourcen-fuer-ihre-lehre/ressourcen-nach-themenbereichen/ki-in-der-hochschullehre.

[15] L. Hoffmann. Chat GPT im Hochschulkontext – eine kommentierte Linksammlung, Hochschulforum Digitalisierung (online): https://hochschulforumdigitalisierung.de/de/blog/chatgpt-im-hochschulkontext-kommentierte-linksammlung.

[16] C. Spannnagel. Regeln für Studierende für den Umgang mit Tools – „Rules for Tools", PH Heidelberg (online): https://csp.uber.space/phhd/rulesfortools.pdf.

[17] J.Gogoll, D.Heckmann, A.Pretschner. Chat GTP und Prüfungsleistungen, FAZ #67, on 20 March 2023, p.18 (in German).

[18] GitMind (online): https://gitmind.com/.

[19] Miro Mind Mapping (online): https://miro.com/mind-mapping/tool/.

[20] A.Kalisz, V.Kalisz et al. OrgPad (online): https://orgpad.com/.

[21] Vít Kalisz, Jiří Kofránek. Univerzální a přesto jednoduchý – OrgPad. Pohledem uživatele a tvůrců, Medsoft 2022 Prague (in Czech, online): https://www.creativeconnections.cz/medsoft/2022/Medsoft_Sbornik_2022_Kalisz.pdf.

[22] Jan Meessen. Wie Künstliche Intelligenz Versicherer revolutioniert, München (Online, in German): https://www.xing.com/news/articles/5606522/paywall/.

[23] Andreas Weck. ChatGPT verfasst mein Anschreiben: So reagieren Personaler auf die KI-Bewerbung (Online, in German: 25.01.2023, 15:55 Uhr): https://www.T3n.de/.

[24] S.B.Bekan. KI wie ChatGPT könnte 300 Millionen Arbeitsplätze beeinträchtigen – einige sind besonders gefährdet (Online, in German: 28.03.2023, 19:32 Uhr): https://www.T3n.de/.

[25] E.Musk. Risiken für die Gesellschaft: Musk und andere Experten fordern Trainingspause für künstliche Intelligenz, LA/ Reuters (Online, in German: 29.03.2023, 11.16 Uhr): https://www.manager-magazin.de/.

[26] S. Russell, D.Dewey, M.Tegmark. Research Priorities for Robust and Beneficial Artificial Intelligence, in "AI Magazine", Winter 2015, pp. 105-114.

[27] Research Priorities for Robust and Beneficial Artificial Intelligence: An Open Letter (Online): https://futureoflife.org/data/documents/research_priorities.pdf.

[28] The Wassenaar Arrangement on Export Controls for Conventional Arms and Dual-Use Goods and Technologies (online): https://www.wassenaar.org/.

[29] MITRE (online): https://www.mitre.org/.

[30] OWASP (online): https://owasp.org/.

[31] SIEM: Security information and event management implementation, NY: McGraw-Hill, 2011, 430p, ISBN 978 007 1701 082.

[32] SIEM / Splunk (online): https://www.splunk.com/.

[33] Central Cooling Plants in Google Datacenters (online): https://www.google.com/.

[34] Marius Feldmann, C&H Dresden GmbH (online): https://www.cloudandheat.com/.

[35] HAEC – Highly-Adaptive Energy-Efficient Computing, 2018 (Sonderforschungsbereich 912) (online): https://tu-dresden.de/zih/forschung/projekte/haec/.

Position Papers of the 18th Conference on Computer Science and Intelligence Systems pp. 87–95
DOI: 10.15439/2023F8583
ISSN 2300-5963 ACSIS, Vol. 36

Analysis of the Public Health Service in Bogotá, Colombia: a Study Based on Customer's Complains and Using Unsupervised Learning Algorithms

Sebastian Quinchia-Lobo, Daniela Salazar-González, Daniel Salas-Álvarez,
Rubén Baena-Navarro, Isaac Caicedo-Castro
0000-0002-2491-737X, 0000-0001-9434-2156, 0000-0002-7097-7883
0000-0001-5055-6515, 0000-0002-7567-3774
Universidad de Córdoba, Socrates Research Team, Faculty of Engineering,
Carrera 6 No. 76-103, 230002 Montería, Córdoba, Colombia
Email: {squinchia31, dsalazargonzalez, danielsalas, rbaena, isacaic}@correo.unicordoba.edu.co

***Abstract*—In this study, our aim is to analyze the public health services in the city of Bogota, Colombia. We used unsupervised learning algorithms for clustering requests, complaints, claims, and denunciations issued to Supersalud in 2021. We collected the data from Supersalud's databases. We adopted clustering algorithms such as K-Means, Bisecting K-Means, and Gaussian Mixture, thus, we evaluated the quality of the combination using the silhouette coefficient. The algorithm with the best clustering quality to generate the clusters has been improved. Of the eight clusters, the first two present the highest incidences, with 181 and 249 affiliates affected for every 2,000 in the year 2021. In the first cluster, with 55% support and 100% confidence, a strong association was found between problems related to medical care facilities and restricted access to health services. In addition, in these two clusters RCCD with pathologies such as chronic communicable and non-communicable diseases (respiratory, diabetes, renal, risk factors and cardiovascular) associated with restricted access to health services were found. In conclusion, The unsupervised grouping allowed to analyze the public health services from the perspective of the RCCD, providing valuable information on the experiences of the users and the challenges in the provision of health services in Bogotá, these findings demonstrate the restriction in the access to health services from different perspectives of a deficient state regarding the provision of health services in the city of Bogotá, Colombia.**

***Index Terms*—Health Indicators, Health Services, Health Statistics, Cluster Analysis, Machine Learning.**

I. Introduction

IN Colombia, health services are regulated by the National Policy for the Provision of Health Services as per Law 1753 of 2015, in conjunction with the surveillance and control body for health services stipulated by Law 1122 of 2007, known as the National Superintendency of Health (Supersalud). Requests, Complaints, Claims, and Denunciations (RCCD) serve as pivotal tools for citizens to voice their concerns to public entities, being recognized as one of the most robust indicators of care quality. However, the quality of service provision in the public health system has been compromised due to adverse situations [1] that impact the contribution of Health Service Providers (HSP) to society.

This work was funded by the University of Córdoba in Colombia

Decision-making in public health can be jeopardized due to biases in the due to the lack of analysis, interpretation and erroneous registration of the same [2], a crucial aspect that can obscure judgments about potential outbreaks and pandemics like Covid-19. In Colombia, restricted access to databases poses a challenge to promoting national public health research [3]. Additionally, the Territorial Health Plan of Bogota 2020-2024 [4] proposes measures to enhance the efficiency of health teams and transition towards a systematized and automated health system.

This study is an extension of a previous investigation [5], where we aim to expand this work with a larger dataset by implementing unsupervised clustering algorithms to group the RCCD filed by healthcare users with Supersalud in Bogota in 2021 [6] and analyze the public healthcare services in this city. Through machine learning, we have utilized open-access data and conducted analyses to meet the primary objective. Our findings suggest that a group of RCCD with chronic pathologies has a strong association with restricted access to health services, as do pathologies such as Covid-19 and cancer.

The study is divided into three parts: the first introduces the challenges of the health sector in Colombia and briefly describes the study conducted along with the methods and techniques applied; the second addresses the findings and the achievement of the objective; and finally, a conclusion is provided that summarizes our findings and discusses their implications for future research.

II. Methodology

This study is classified as a quantitative research of descriptive type, adapted from other studies [7], widely explained by Sampieri [8], with an extended data mining approach of the Cross-Standard Process for Data Mining in Industry (CRISP-DM) [9] to analyze RCCD and meet the objective of the study. The phases of the process used are described below:

©2023, PTI

Topical area: Advanced Artificial Intelligence in Applications

A. Software

For the analysis of information, the Python 3.11.1 programming language, the Apache Spark platform and the PySpark 3.3.2 library were used. These tools allow large amounts of information to be processed in a distributed, parallel, and replicated manner [10], which guarantees reliable and reproducible results. In addition, the Google Colab development environment was used, which makes it easy to work with Jupyter Notebooks online.

B. Data

The dataset for this study is sourced from the open RCCD database interposed to Health Promoting Entities (HPE) with Supersalud in Colombia for the year 2021 [6]. Acquired on February 15, 2023, the data adheres to the *Habeas Data* Law, ensuring obfuscation to protect individual identities. Each RCCD is assumed to represent a unique individual. Our analysis is centered on Bogotá's RCCD, encompassing both the contributory and subsidized regimes. Nationally, 993,349 RCCDs were recorded, with Bogotá being the city with the most RCCD records, accounting for a total of 226,230. Specifically, the contributory and subsidized regimes in Bogotá accounted for 213,375 RCCDs out of 7,927,520 affiliates as of December 2021 [11]. This study narrows its focus to these regimes in Bogotá, as they comprise 94% of the city's RCCDs, split between 169,431 (contributory) and 43,944 (subsidized).

1) Understanding the problem and its data: To comprehend the problem and its data, we integrated the business understanding and data comprehension phases of CRISP-DM. Initial assessments confirmed the integrity of two key features: the municipality code of the entity receiving the RCCD and the affiliation regime of the affected individual. Filtering was applied based on these features, specifically `ENT_COD_MPIO` (municipality code, with 11001 representing Bogotá) and `AFEC_REGAFILIACION` (affiliation regime, focusing on subsidized and contributory). A subsequent deep dive into RCCD features revealed several categorical attributes crucial for understanding sector challenges, including gender, macro-motive, life risk, age range, pathology, and high cost. These attributes were found to be comprehensive, with no data loss observed. Their value domains after applying the filter are distributed as follows:

- **The gender** named in the data set as `AFEC_GENERO` determines the gender of the affected person. With a cardinality of 2, it takes two values: Man and Woman, at the national level (without applying the filter) another value called "Not Applicable" was observed.
- **The life risk** named in the data set as `RIESGO_VIDA` determines if the reason for which the RCCD was filed puts the life of the affected person at risk. With a cardinality of 2, it takes two possible values Yes and No (Without applying the filter, no others were observed).
- **The age range** of the affected person in the data set called `AFEC_EDADR`. With a cardinality of 9, due to the obfuscation of the data, it takes values such as: Between 0 and 5 years, between 6 and 12 years, between 13 and 17 years, between 18 and 24 years, between 25 and 29 years, between 30 and 37 years, between 38 and 49 years, between 50 and 62 years and over 63 years, at the national level (without applying the filter) another value called "Not Applicable" was observed.
- **The macro-motive** named in the data set as `MACROMOTIVO`. With a cardinality of 6, it takes the following categorical values (Without applying the filter, no others were observed): Restriction in access to health services, Deficiency in the effectiveness of health care, User dissatisfaction with the administrative process, Non-recognition of economic benefits, Lack of availability or inappropriate management of human and physical resources for care and Petitions, complaints and claims filed by HPS-HPE, territorial entities and control and surveillance agencies.
- **The high cost** feature named in the data set as `ALTO_COSTO`, this determines if the reason for the RCCD is related to a high cost disease, it presents a cardinality of 22 and takes the following categorical values (Without applying the filter, it does not others were observed): Not Applicable, Peritoneal dialysis, Hemodialysis, Management of patients in intensive care units, Diagnosis and management of the HIV-infected patient, Chemotherapy and radiation therapy for cancer, Medical-surgical management of major burns, Management of major trauma among others.
- **The pathology** feature named in the data set as `PATOLOGIA_1`, contains the disease related to the reason that the RCCD occurred. This feature have a cardinality of 22 and take the following categorical values (Without applying the filter, no others were observed): Chronic non-communicable respiratory diseases, Problems related to health care facilities or other health services, Non-communicable chronic diseases - diabetes, Osteoarticular diseases, Cancer, Chronic non-communicable cardiovascular diseases among others.

2) Data preparation and experimental configuration: At this stage, the CRIPS-DM data preparation phase was adopted and an experimental setup was included. Regarding the preparation of the data, the following considerations were taken into account:

- As can be seen previously, the "high cost" feature, which determines the cause of a high value for which the RCCD is interposed, has a high cardinality, for this reason it was reduced to using two values yes and no, in the case if it does not have any high cost (Not Applicable) its value is no and yes in otherwise.
- The pathology called "problems related to medical care facilities or other health services", despite not being its own pathology, was not excluded because it facilitates understanding of the impact and demonstrates shortcomings in the quality of health care provision services.

For the experimental configuration, combinations of features were created according to the objective of the study for the

subsequent phases (see table I).

We use one-hot coding [12] because the domain of the study features is categorical (see section II-B1). This coding assumes a categorical variable d, which takes values in the set $O = o_1, o_2, \ldots, o_H$, one-hot transforms d into an H-dimensional vector p, such that each dimension h_i comprises a value between zero and one corresponding to its value in the set O (see eq. 1). This process increases the dimensionality of the database, which implies a slightly higher processing cost due to this increase.

$$h_i = \begin{cases} 1 \text{ if } d = o_i \\ 0 \text{ if } d \neq o_i \end{cases} \tag{1}$$

Where h_i is the ith dimension of a categorical variable d, which takes values in the set O. o_i is the ith value taken by the categorical variable d.

Utilizing one-hot, categorical values like "Over 63 Years" in age range are transformed into binary columns. If an RCCD instance has this value, it's marked as 1; otherwise, it's 0. Because we use Apache Spark, we use their optimized approach that customizes one-hot for large data sets that creates a single vector column, capturing all values per feature (see eq. 1).

3) Clustering Algorithms: In this phase, the CRISP-DM modeling phase was adopted, using the previously prepared data set. Accordingly, it is proposed to use the unsupervised grouping technique to group the RCCD according to the study variables. This would allow segmenting the experiences of patients. For example: if a cluster shows that a group of RCCD have restrictions regarding some pathologies, the regulatory entities of the health sector could investigate ways to reduce the restrictions for these patients.

In addition to the above, due to the nature of the data, they do not have a predefined class or category that allows separating the data to predict, classify or group [13], therefore, in this research, we have adopted unsupervised machine learning clustering algorithms, such as K-Means, Bisecting K-Means, Gaussian Mixture [14] from the Apache Spark Pyspark library.

K-Means is an unsupervised clustering algorithm guided by proximity in a search space, which is calculated in closeness to the centroids (central point of a cluster that defines it) [15] until it is minimal, noting that it considers each cluster as convex due to the Euclidean distance determining the closeness to the centroids (see eq. 2) [16]. The implementation of K-Means available in the PySpark library for distributed processing of large volumes of information of Apache Spark in its official documentation [17] mentions that it uses a variant of the original algorithm called K-Means++, in whose particular case according to the documentation is based on the article by Bahmani [18], where a controlled random sample is used (see eq. 3), generating distant centroids. The following definitions have been recovered from the previously mentioned article:

$$d(x, C) = \min_{c \in C} ||x - c||^2 \tag{2}$$

$$C \leftarrow C \cup \{x\}; x \in X \text{ with a probability of } \frac{d^2(x, C)}{\Phi_X(C)} \tag{3}$$

Where X is the set of observations in the dataset, x is a specific observation of X, C is the group of selected centroids, c is a center of the set C, $d(x, C)$ is the Euclidean distance from x to the closest centroid in C, $\Phi_X(C)$ is the objective function to minimize using the Euclidean distance from the points in X to the centroids in C.

Bisecting K-Means is a variant of K-Means as its name indicates, the difference from the original being that it splits the data into subgroups forming a tree as its nodes subdivide until they are indivisible, in such a way that the specified or identified k clusters are generated, using K-Means with its Euclidean distance (see eq. 2). According to the official documentation of the implemented library [19], it is based on the article by Steinbach [20, 21], whose implementation is still used with some updates.

Gaussian Mixture is a probability-based algorithm which allows generating clusters considering that each of them follows a Gaussian distribution. According to the implemented library, the expectation maximization originally described by Dempster [22] is used to generate Gaussian Mixture Models (GMMs) [23]. These models are given by the multivariate probability density function, see eq. 4 in [24].

$$g(x, \Phi) = \sum_{k=1}^{K} P_k f(x, \mu_k, \Sigma_k) \tag{4}$$

4) Evaluation and quality metrics: This phase extends from the evaluation phase of CRISP-DM, in which an analysis was carried out using the silhouette coefficient as a criterion to select the clustering algorithms and determine their optimal configuration in terms of number of clusters and experimental parameters (see table I), where the different configurations were evaluated, for each one a range of groups between two and eight was assigned due to the highest silhouette coefficient being obtained with eight clusters.

Without classes, the evaluation and comparison between these algorithms is challenging, where it is necessary to have alternatives to similarity metrics between clusters, such as the silhouette coefficient, which is a dimensionless measure [25] for determining clustering coherence [26], ensuring optimal quality, with values ranging between -1 and 1, 1 being the best result and -1 the opposite, see eq. 5

$$s(i) = \frac{b(i) - a(i)}{\{a(i), b(i)\}}, \text{ where } -1 \leq s(i) \leq 1 \tag{5}$$

Where i is the ith observation, $s(i)$ is the silhouette coefficient for observation i, $a(i)$ is the average distance between i and all other observations in the same cluster, $b(i)$ are all the observations of the nearest cluster different from i.

5) Analysis tools: To understand the magnitude of the problem, it is necessary to determine the proportional (cumulative) incidence of each cluster [27, 28], taking as parameters the total number of affiliates in December 2021 of the regimes of this investigation extracted from external reports [11], using as a reference the name and code of the municipality (Bogotá, 001) and the number of RCCD for each cluster, in terms of

TABLE I
COMBINATIONS OF FEATURES USED TO PERFORM TESTS ON THE DATASET.

Feature Combination Identifier	Features					
	Gender	macro-motive	Life risk	Age range	Pathology	High cost
1	x		x	x		
2	x	x	x	x		
3	x	x	x		x	
4	x	x	x	x	x	
5	x	x	x		x	x
6	x	x	x	x	x	x
7			x		x	x
8		x	x		x	x
9		x	x			x

every 2,000 affiliates per year (see eq. 6):

$$I_p(k) = \frac{w_k}{T_a} p \tag{6}$$

Let k be the kth cluster, $I_p(k)$ the proportional incidence of k, w_k the number of observations in k, T_a the total number of affiliates in the regimes of study, and p the estimation proportion in terms of affiliates per year.

It is also necessary to use association rules to evaluate the frequency and probability of occurrence of a set of values in each cluster, for which support and confidence are used, which through the following expressions recovered from external sources [29]:

$$\text{Support} = P(T \cap K) \tag{7}$$

$$\text{Confidence} = P(K|T) = \frac{P(T \cap K)}{P(T)} \tag{8}$$

Let T and K be two set of items.

III. RESULTS

A. Clustering Outcomes

After carrying out several experiments, the best experiments were selected based on their silhouette coefficients, in table II where the 10 best results are presented, ordered from highest to lowest, of these 10 best results they have the same combination of features given by macro-motive, life risk and high cost, identified with ID 9 (see table I). The two best results, with eight and seven clusters respectively, were obtained using the K-Means algorithm and presented an acceptable adjustment time.

Figure 1 depicts heat maps using the data from the 189 iterations performed for each algorithm using the number of clusters, features combinations, and their silhouette coefficient. It can be observed that the highest silhouette coefficient is found in combinations 7 and 9 of the clustering algorithms. Additionally, it is highlighted that K-Means achieved a better silhouette coefficient in the conducted tests, while Gaussian Mixture exhibited negative silhouette coefficient and other very low values, indicating its limitations.

Figure 2 shows the processing cost (time it takes each algorithm to process the data) for the three clustering algorithms: K-Means, GaussianMixture and BisectingKMeans. It can be observed that K-Means obtained the lowest processing cost, with an interquartile range (IQR) of 2.5 seconds and a total range of 7.05 seconds, with a median of 8.76 seconds, a mean of 9.2 seconds, and an outlier of 29.57 seconds. Unlike the other two algorithms, whose performance was lower due to having a longer processing time. In GaussianMixture a greater dispersion is observed, with a IQR of 27.23 seconds and a total range of 70.6 seconds, with a median of 19.32 seconds and a mean of 28.38 seconds. This algorithm has 2 outliers of 80.77 and 103.84 seconds. BisectingKMeans has a median of 41.59 seconds and a mean of 38.2 seconds, also an IQR of 15.82 seconds and a total range of 34.11 seconds. This algorithm has no outliers.

B. Identified Clusters

In summary, with the previous findings, the best result was obtained when using K-Means with the combinations of features given by macro-motive, life risk and high cost, identified with ID 9 (see table I) and with 8 clusters compared to the others through the silhouette coefficient; In a second run using these same parameters, eight clusters with a silhouette coefficient of approximately 0.97 were identified. The RCCD instances with the identified clusters were matched to the prepared data to be able to use the other features such as pathology, gender, and age range, which were not present in the combination of features (experimental setup) used for clustering.

For each cluster, the cumulative incidence was calculated in terms of 2,000 affiliates in December 2021 for each cluster (see eq 6), two of them present higher incidences compared to the total corresponding to clusters one and two (see table III), the remaining clusters present low incidences. From Table (see table IV), it can be seen that the amount of RCCD accumulates in some values of the macro-motive feature of the clusters. On the other hand, in most of the clusters, chronic diseases are present.

The first cluster represents 181 out of every 2,000 affiliates in December 2021. This cluster presents a life risk, but not a high cost, with a support $P(X \cap Z) = \{$ Quantity of RCCD in the cluster with chronic communicable diseases and noncommunicable (respiratory, diabetes, renal, risk factors and cardiovascular) that have the macro-motive of restriction in access

TABLE II
ITERATIONS OF THE DIFFERENT COMBINATIONS OF FEATURES, ALGORITHMS AND NUMBER OF CLUSTERS

Algorithm	Number of clusters	Silhouette coefficient	Model fitting time in seconds
KMeans	8	0.96	7.55
KMeans	7	0.95	5.15
GaussianMixture	8	0.93	46.80
BisectingKMeans	8	0.91	47.08
KMeans	6	0.91	7.00
GaussianMixture	7	0.89	19.32
GaussianMixture	6	0.88	36.84
BisectingKMeans	7	0.86	46.10
GaussianMixture	5	0.85	36.32
BisectingKMeans	6	0.83	42.89

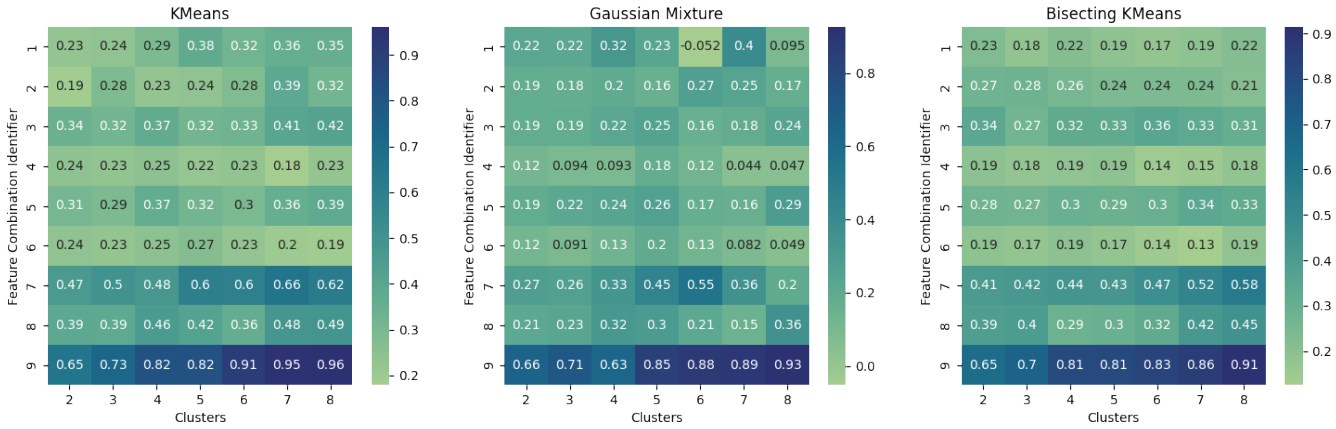


Fig. 1. Heat maps of the clustering algorithms with the combinations of features and the number of clusters.

TABLE III
QUANTITY OF RCCDS RECEIVED FOR EACH VALUE PER FEATURE - PART 1

Feature	Value	Quantity of RCCD per cluster							
		1	2	3	4	5	6	7	8
		size = 71819	size = 98843	size = 6831	size = 7826	size = 8059	size = 15411	size = 3735	size = 851
		$I = 181$	$I = 249$	$I = 17$	$I = 20$	$I = 20$	$I = 39$	$I = 9$	$I = 2$
Range of age	0 - 5	3184	4806	276	196	30	752	153	19
	6 - 12	2562	4068	196	76	31	591	97	5
	13 - 17	2489	2991	151	104	21	411	111	8
	18 - 24	4517	5257	503	241	319	1281	208	43
	25 - 29	5021	6168	581	382	1037	1363	279	55
	30 - 37	8068	10778	948	827	2727	2121	435	90
	38 - 49	11205	15949	1330	1471	2255	2657	570	167
	50 - 62	13262	21064	1236	1994	1174	2847	675	205
	63+	21511	27762	1610	2535	465	3388	1207	259
Gender	Man	30498	38427	2815	4076	3764	6503	1600	432
	Woman	41321	60416	4016	3750	4295	8908	2135	419
High cost	No	71819	98045	6708	0	7898	14923	3735	0
	Yes	0	798	123	7826	161	488	0	851
Risk of life	No	0	98843	6831	0	7847	11849	0	0
	Yes	71819	0	0	7826	212	3562	3735	851

TABLE IV
QUANTITY OF RCCDS RECEIVED FOR EACH VALUE PER FEATURE - PART 2

Feature	Value	Quantity of RCCD per cluster							
		1	2	3	4	5	6	7	8
Macro-motive	Deficiency in the effectiveness of health care	0	0	6831	0	0	0	3735	851
	Lack of availability or inappropriate management of human and physical resources for care	133	119	0	19	1	0	0	0
	User dissatisfaction with the administrative process	0	0	0	0	0	15186	0	0
	Non-recognition of economic benefits	0	0	0	0	8058	0	0	0
	Petitions, complaints and claims filed by HPS-HPE, territorial entities and control and surveillance agencies	336	0	0	86	0	225	0	0
	Restriction in access to health services	71350	98724	0	7721	0	0	0	0
Pathology	Covid-19	13285	12	9	452	43	482	584	48
	Intensive care for any pathology	3	21	2	43	0	10	0	12
	Cancer	9131	441	159	3183	231	640	710	479
	Overall effectiveness of care	15	8	6	2	0	1	3	0
	Chronic communicable disease	37	8	1	1	1	1	0	0
	Vector-borne disease	3	1	1	0	0	0	1	0
	Chronic non-communicable respiratory diseases	6196	1566	170	166	84	500	384	12
	Chronic non-communicable diseases - diabetes	3291	3319	158	134	40	332	131	6
	Chronic non-communicable diseases - renal	954	471	55	844	77	142	47	68
	Chronic non-communicable diseases - risk factors	588	1807	200	22	34	190	37	3
	Chronic non-communicable cardiovascular diseases	9900	13473	667	664	128	1410	431	30
	Orphan diseases	1982	181	28	61	15	95	120	5
	Immune-preventable diseases	6	10	0	1	0	0	0	0
	Neurological diseases	2378	718	31	14	23	152	62	2
	Osteoarticular diseases	4266	14809	757	243	413	1137	161	9
	Great burn	6	3	2	18	1	4	0	4
	Maternal-child	3352	3660	246	119	45	739	255	7
	Not applicable	134	88	12	2	17	38	1	0
	Problems related to medical care facilities or other health services	10209	54709	3932	360	6797	8742	489	20
	Prosthetic hip or knee joint replacements	20	130	4	274	8	19	1	6
	Mental health	5831	1118	168	33	58	476	306	3
	Oral health	60	2158	198	2	24	188	1	1
	Sexual and reproductive health	10	57	4	0	0	8	2	0
	HIV AIDS and other sexually transmitted diseases	162	75	21	1188	20	105	9	136

to health services } / {cluster size} $= 29\%$ and a confidence $(X \rightarrow Z) = P(Z|X)$ = { Quantity of RCCD in the cluster with communicable and non-communicable chronic diseases (respiratory, diabetes, renal, risk factors and cardiovascular) that have the macro-motive of restriction in access to health services} / { Quantity of RCCD in the cluster with chronic communicable and non-communicable diseases (Respiratory, diabetes, renal, risk factors and cardiovascular)} $= 99\%$, this suggests a pattern in the data from this cluster. The pathologies of covid-19 and Cancer are also highlighted, also with the previous macro-motive, which have a support $P(X, Z) =$ { Quantity of RCCD in the cluster with covid-19 and cancer that have the macro-motive of restriction in access to health services } / { cluster size} $= 31\%$ and with a confidence$(X \rightarrow Z) = P(Z|X)$ = { Quantity of RCCD in the cluster with covid-19 and cancer who have the macro-motive for restriction of access to health services} / { Quantity of RCCD in the cluster with covid-19 and cancer } $= 99\%$, this reveals a strong pattern in the data, something that is worrisome since it reveals that in this cluster to these pathologies, access to health services is restricted, which can increase their risk of serious complications or long-term effects by not having timely or

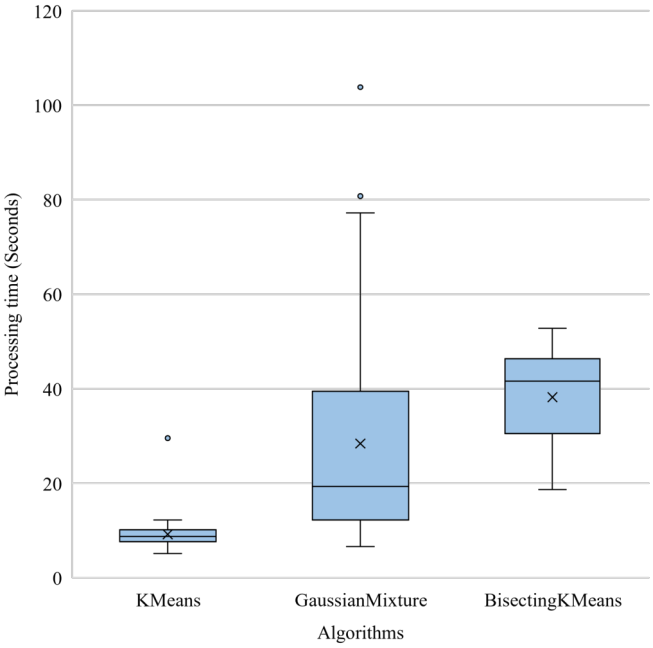


Fig. 2. Boxplot of the processing time of the different algorithms

adequate treatment; In addition, cancer and Covid-19 are two pathologies to highlight because cancer patients usually contract an immunosuppressed state that increases the risk of infections and exposes them to complications [30], both pathologies according to the previous results are restricted to them. access to health services, this causes greater concern, because health should be for everyone and should not be a privilege.

In the second cluster, 61% correspond to women, without a life risk and for the most part does not present a high cost, this represents 249 of every 2,000 affiliates in December 2021. With a support $P(X \cap Z) = \{$ Quantity of RCCD in the cluster with the pathology called problems related to medical care facilities or other health services that have the macro-motive of restriction in access to health services$\}$ / {Cluster size$\}$ $= 55\%$ and a confidence $(X \rightarrow Z) = P(Z|X) = \{$ Quantity of RCCD in the cluster with the pathology called problems related to health care facilities or other health services that have the macro-motive of restriction in access to services of health$\}$ / $\{$ Quantity of RCCD in the cluster with the pathology called problems related to medical care facilities or other health services$\}$ $= 100\%$, this indicates a strong pattern in the data, revealing the shortcomings of the users in the public health care, which are not satisfied with the care and has a strong association with the restriction of health services. It is also highlighted from this cluster that with a support $P(X \cap Z) = \{$ Quantity of RCCD in the cluster with chronic communicable and non-communicable diseases (Respiratory, diabetes, renal, risk factors and cardiovascular) that have the macro-motive of restriction in access to health services $\}$ / {Cluster size$\}$ $= 21\%$ and confidence $(X \rightarrow Z) = P(Z|X) = \{$ Quantity of RCCD in the cluster with chronic communicable and non-communicable diseases (Respiratory, diabetes, renal, risk factors and cardiovascular) that have the macro-motive of restriction in access to health services$\}$ / $\{$ Quantity of RCCD in the cluster with communicable and non-communicable chronic diseases (Respiratory, diabetes, renal, of risk and cardiovascular)$\}$ $= 100\%$, this demonstrates the strong association with this macro-motive, in addition to these chronic pathologies that can affect the quality of life of the affected person if the restriction in access to health services persists, implying a imminent risk to public health, due to the fact that they are not given sufficient attention or the state of medical care services are not suitable, leading to a worsening of their health status, such that it presents a life risk to this population.

IV. Discussion

In our study, we employed unsupervised clustering algorithms, including K-Means, Bisecting K-Means, and Gaussian Mixture, to analyze the Right to Health Care Claims (RCCD) in Bogotá, Colombia's public health services. Our findings revealed that K-Means outperformed the other two algorithms in terms of efficiency, enabling us to segment user experiences and gain a nuanced understanding of the challenges they encounter when accessing health services.

Our analysis underscores that health services for chronic pathologies, predominantly evident in clusters 1 and 2, are notably constrained. Within these clusters, a recurring macro-motive consistently yielded the highest RCCD count. Notably, cluster 1 emphasized issues within medical care facilities. These observations resonate with Lemy's research [30], which identified administrative barriers in accessing health services. Furthermore, the prominence of Covid-19 and Cancer pathologies in the first cluster indicates restricted health service access during the 2021 pandemic. This aligns with other studies [31] that observed a decline in preventive services and an absence of comprehensive vaccination strategies for Covid-19, especially for cancer patients.

Interestingly, while some rules exhibited low support, they demonstrated high confidence, indicating robust relationships between specific variables within data subsets. For instance, a rule associating Covid-19 and cancer pathologies with restricted health service access exhibited 31% support but a staggering 99% confidence. Such patterns can be instrumental in discerning trends within specific data groups. However, it's pivotal to remember that correlation doesn't equate to causation, necessitating further research to ascertain any causal links.

The implications of our findings are profound. They accentuate the imperative to enhance the health system's efficacy and structure, especially in crisis scenarios like pandemics. We advocate for the relevant authorities to devise strategies bolstering the availability and accessibility of public health care services, drawing insights from our study.

However, our research is not without limitations. The scope was confined to RCCD in Bogotá, covering the contributory and subsidized healthcare regimes, the two regimes with the

most RCCD. Our quantitative approach, based on RCCD from Supersalud, might not capture the data's full depth. The study's lens was solely on user perception and satisfaction, overlooking other potential influencers like healthcare personnel, infrastructure, or funding. We experimented with various feature combinations for the clustering algorithms, but other configurations remain unexplored. For instance, we limited our cluster count to eight based on the optimal coefficient. Lastly, due to data obfuscation for legal compliance, individual identification was impossible, leading us to assume each RCCD represents a distinct individual in our analysis.

V. Conclusion

This study underscores the power of sophisticated data analysis techniques in the domain of public health, specifically when working with RCCD. By leveraging mathematical and statistical models for information preprocessing and employing algorithms like K-Means, Bisecting K-Means, and Gaussian Mixture, we achieved a comprehensive analysis of public health services in Bogotá, Colombia. The silhouette coefficient played a pivotal role, ensuring the best clustering quality and preventing the generation of ambiguous or non-informative results. With a strong association identified, having a support of 55% and a confidence of 100%, we found significant issues related to healthcare facilities and restrictions in access to health services.

One of the standout findings was the evident deficiencies in the health system, especially concerning the quality of services. Chronic pathologies, both communicable and non-communicable, were prominently present in clusters 1 and 2, with the highest incidence. These clusters revealed a significant restriction in access to health services, with this restriction being a dominant motive. The macro-motives in these two clusters reflected the challenges in accessing health services, emphasizing the urgency to prioritize and address these issues. Without timely and appropriate care, the risk to patients' lives is substantially heightened, especially given the current restrictions in the public health sector.

Considering these findings, it's recommended to expand the scope of this study, incorporating data from different regions of Colombia to capture a more comprehensive national perspective. Such an approach could unveil crucial insights into the challenges faced by public health services across the country. Moreover, a deeper exploration of the patterns and strong associations identified in this study is essential. Understanding the intricate relationships between various variables and their impact on health service quality can pave the way for more informed administrative decisions, ultimately enhancing patient care in the health sector.

Acknowledgment

We thank Universidad de Córdoba in Colombia for supporting this study and Supersalud for publishing the dataset used in this study. Too thanks to the projects SFCB-01-21 and FI-01-22 of the Universidad de Cordoba. Caicedo-Castro thanks the Lord Jesus Christ for blessing this project. Finally, we thank the anonymous reviewers for their comments that contributed to improve the quality of this article

References

[1] D. Mendieta and G. Rojas, "Corruption the biggest epidemic that colombia suffers," *Revista Opiniao Juridica*, vol. 19, no. 32, pp. 296–315, Sep. 2021. [Online]. Available: https://periodicos.unichristus.edu.br/opiniaojuridica/article/view/3979

[2] S. C. Johnson, M. Cunningham, I. N. Dippenaar *et al.*, "Public health utility of cause of death data: applying empirical algorithms to improve data quality," *BMC Medical Informatics and Decision Making*, vol. 21, no. 175, p. 20, Dec. 2021. [Online]. Available: https://bmcmedinformdecismak.biomedcentral.com/articles/10.1186/s12911-021-01501-1

[3] J.-S. Franco and D. Vizcaya, "Availability of secondary healthcare data for conducting pharmacoepidemiology studies in Colombia: A systematic review," *Pharmacology Research & Perspectives*, vol. 8, Oct. 2020. [Online]. Available: https://onlinelibrary.wiley.com/doi/10.1002/prp2.661

[4] Secretaría Distrital de Salud de Bogotá D.C., "Plan Territorial de Salud para Bogotá D.C. 2020-2024," Secretaría Distrital de Salud de Bogotá D.C., Bogotá D.C., Tech. Rep., 2020. [Online]. Available: https://subredsuroccidente.gov.co/planeacion/DOCUMENTO%20PTS%202020-2024%20%2027042020.pdf

[5] S. Quinchia-Lobo and D. Salazar-González, "Análisis exploratorio de las pqrd del sector salud mediante aprendizaje no supervisado para identificar las principales barreras y oportunidades de mejora en la prestación del servicio en la salud pública del municipio de montería," B.Sc. thesis, Universidad de Córdoba, Monteria, Córdoba, Jul. 2023, supervisors: Salas-Alvarez D. and Baena-Navarro R. [Online]. Available: https://repositorio.unicordoba.edu.co/handle/ucordoba/7408

[6] SUPERSALUD. Base de datos pqrd 2021 - csv — portal de datos abiertos de la sns. [Online]. Available: https://mapas.supersalud.gov.co/arcgisportal/apps/sites/#/datos-abiertos/datasets/3824e636c1b748269364c0e57c680d58/about

[7] M. Hinojosa, I. Derpich, M. Alfaro *et al.*, "Procedimiento de agrupación de estudiantes según riesgo de abandono para mejorar la gestión estudiantil en educación superior," *Texto Livre*, vol. 15, p. 22, Mar. 2022. [Online]. Available: https://periodicos.ufmg.br/index.php/textolivre/article/view/37275

[8] R. Hernández Sampieri, C. Fernández Collado, and P. Baptista Lucio, *Metodología de la investigación*, 5th ed. México, D.F: McGraw-Hill, 2010.

[9] C. Schröer, F. Kruse, and J. M. Gómez, "A Systematic Literature Review on Applying CRISP-DM Process Model," *Procedia Computer Science*, vol. 181, pp. 526–534, 2021. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S1877050921002416

[10] E. Nazari, M. H. Shahriari, and H. Tabesh, "BigData Analysis in Healthcare: Apache Hadoop , Apache spark and Apache Flink," *Frontiers in Health Informatics*, vol. 8, no. 1, pp. 92–101, Jul. 2019. [Online]. Available: http://ijmi.ir/index.php/IJMI/article/view/180

[11] ADRES. (2022) Reporte de afiliados por departamento y municipio. [Online]. Available: https://www.adres.gov.co/eps/bdua/Paginas/reporte-afiliados-por-departamento-y-municipio.aspx

[12] M. K. Dahouda and I. Joe, "A Deep-Learned Embedding Technique for Categorical Features Encoding," *IEEE Access*, vol. 4, p. 12, 2016.

[13] W. Bao, N. Lianju, and K. Yue, "Integration of unsupervised and supervised machine learning algorithms for credit risk assessment," *Expert Systems with Applications*, vol. 128, pp. 301–315, Aug. 2019. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0957417419301472

[14] U. N. Wisesty and T. R. Mengko, "Comparison of dimensionality reduction and clustering methods for sars-cov-2 genome," *Bulletin of Electrical Engineering and Informatics*, vol. 10, no. 4, pp. 2170–2180, 2021. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-85111115089&doi=10.11591%2fEEI.V10I4.2803&partnerID=40&md5=9f4a6b2b087f1e835402560d8081947c

[15] S. Jian, D. Li, and Y. Yu, "Research on Taxi Operation Characteristics by Improved DBSCAN Density Clustering Algorithm and K-means Clustering Algorithm," *Journal of Physics: Conference Series*, vol. 1952, no. 4, p. 7, Jun. 2021. [Online]. Available: https://iopscience.iop.org/article/10.1088/1742-6596/1952/4/042103

[16] K. P. Sinaga and M.-S. Yang, "Unsupervised K-Means Clustering Algorithm," *IEEE Access*, vol. 8, pp. 80 716–80 727, 2020. [Online]. Available: https://ieeexplore.ieee.org/document/9072123/

[17] Apache Software Foundation. Kmeans — pyspark 3.3.2 documentation. [Online]. Available: https://spark.apache.org/docs/latest/api/python/reference/api/pyspark.ml.clustering.KMeans.html#pyspark.ml.clustering.KMeans

[18] B. Bahmani, B. Moseley, A. Vattani *et al.*, "Scalable k-means++," *Proceedings of the VLDB Endowment*, vol. 5, no. 7, pp. 622–633, Mar. 2012. [Online]. Available: https://dl.acm.org/doi/10.14778/2180912.2180915

[19] Apache Software Foundation. Bisectingkmeans — pyspark 3.3.2 documentation. [Online]. Available: https://spark.apache.org/docs/latest/api/python/reference/api/pyspark.ml.clustering.BisectingKMeans.html#pyspark.ml.clustering.BisectingKMeans

[20] M. Steinbach, G. Karypis, and V. Kumar, "A comparison of document clustering techniques," *KDD Workshop on Text Mining*, 2000.

[21] M. Vichi, C. Cavicchia, and P. J. F. Groenen, "Hierarchical Means Clustering," *Journal of Classification*, vol. 39, no. 3, pp. 553–577, Nov. 2022. [Online]. Available: https://link.springer.com/10.1007/s00357-022-09419-7

[22] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum Likelihood from Incomplete Data Via the *EM* Algorithm," *Journal of the Royal Statistical Society: Series B*, vol. 39, no. 1, pp. 1–22, Sep. 1977. [Online]. Available: https://onlinelibrary.wiley.com/doi/10.1111/j.2517-6161.1977.tb01600.x

[23] Apache Software Foundation. Gaussianmixture — pyspark 3.3.2 documentation. [Online]. Available: https://spark.apache.org/docs/latest/api/python/reference/api/pyspark.ml.clustering.GaussianMixture.html#pyspark.ml.clustering.GaussianMixture

[24] K. Aziz, D. Zaidouni, and M. Bellafkih, "Leveraging resource management for efficient performance of Apache Spark," *Journal of Big Data*, vol. 6, p. 23, Dec. 2019. [Online]. Available: https://journalofbigdata.springeropen.com/articles/10.1186/s40537-019-0240-1

[25] P. J. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," *Journal of Computational and Applied Mathematics*, vol. 20, pp. 53–65, Nov. 1987. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/0377042787901257

[26] K. R. Shahapure and C. Nicholas, "Cluster Quality Analysis Using Silhouette Score," in *2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA)*. sydney, Australia: IEEE, Oct. 2020, pp. 747–748. [Online]. Available: https://ieeexplore.ieee.org/document/9260048/

[27] A. Fajardo-Gutiérrez, "Medición en epidemiología: prevalencia, incidencia, riesgo, medidas de impacto," *Revista Alergia México*, vol. 64, no. 1, pp. 109–120, Feb. 2017. [Online]. Available: http://revistaalergia.mx/ojs/index.php/ram/article/view/252

[28] L. Rychetnik, P. Hawe, E. Waters *et al.*, "A glossary for evidence based public health," *Journal of Epidemiology & Community Health*, vol. 58, pp. 538–545, 2004. [Online]. Available: https://jech.bmj.com/lookup/doi/10.1136/jech.2003.011585

[29] D. T. Larose and C. D. Larose, *Discovering Knowledge in Data: An Introduction to Data Mining*. Hoboken, NJ, USA: John Wiley & Sons, Inc., 2014. [Online]. Available: http://doi.wiley.com/10.1002/9781118874059

[30] O. M. Al-Quteimat and A. M. Amer, "The Impact of the COVID-19 Pandemic on Cancer Patients," *American Journal of Clinical Oncology*, vol. 43, no. 6, pp. 452–455, Jun. 2020. [Online]. Available: https://journals.lww.com/10.1097/COC.0000000000000712

[31] L. Bran Piedrahita, A. Valencia Arias, L. Palacios Moya *et al.*, "Barreras de acceso del sistema de salud colombiano en zonas rurales: percepciones de usuarios del régimen subsidiado," *Hacia la Promoción de la Salud*, vol. 25, no. 2, pp. 29–38, Jul. 2020. [Online]. Available: https://revistasojs.ucaldas.edu.co/index.php/hacialapromociondelasalud/article/view/2358

Position Papers of the 18th Conference on Computer Science and Intelligence Systems pp. 97–102
DOI: 10.15439/2023F4728
ISSN 2300-5963 ACSIS, Vol. 36

Multi-Criteria Decision-Making by Approximation in the Domain of Linguistic Values

Leszek Rolka
0000-0003-0083-8893
Rzeszów University of Technology
Al. Powstańców Warszawy 8, 35-959 Rzeszów, Poland
Email: leszekr@prz.edu.pl

***Abstract*—This paper presents a method of determining the set of alternatives, with respect to a subset of fuzzy criteria, that have the greatest degree of accordance with preferences given by a decision-maker. We apply two kinds of alternatives in the process of decision-making. The candidate solutions will be selected from a (large) universe of (real) alternatives. Their membership degrees in the linguistic values of all fuzzy criteria is assigned by an expert. A set of (imaginary) reference alternatives is generated to express the expectations of the decision-maker, who assigns membership degrees in the most preferred linguistic values of fuzzy criteria. We define the notion of approximation of a reference alternative by a real alternative in the domain of linguistic values of criteria, and introduce a measure of accordance of the real alternative with the reference alternative.**

I. Introduction

MODELLING and analysis of the decision-making activity preformed by humans strongly depends on the quality of data obtained from the real-world reasoning processes. The rough set theory [1] is a paradigm aimed at discovering uncertainty, inconsistency, and redundancy that can be often found in information systems. It could be successfully combined with different ways of knowledge representation in the form of hybrid approaches, such as fuzzy-rough decision models [2]–[6]. Both the rough set models and the fuzzy knowledge representation should be utilized in attempts to create systems that could be helpful in finding an optimal alternative in complex situations. It is especially important in the case of huge information systems, with a vast number of alternatives that are characterized by many criteria.

Even a skilled expert can hardly manage a difficult decision-making task, not only due to the size of the problem, but also because of natural contradictions between various criteria. Furthermore, when a group of experts is involved in finding an optimal decision, we can expect solutions that may be inconsistent. It is obvious that the human experts need to be supported in solving both the multi-criteria and the group decision-making tasks. The most popular algorithms applied to this end are SAW, TOPSIS, and AHP [7], [8].

In order to model actual decision-making processes, it is also necessary to admit subjectivity in goal functions and constrains. In a real-world situation, one cannot be restricted to objective (numerical) criteria only, such as cost or benefit, but has also to take into account subjective (vague) criteria. Therefore, we can observe that the standard (crisp-oriented) methods were adopted to work with fuzzy numbers and linguistic values in the multi-criteria [9]–[11], and in the group decision-making [12]–[15].

Motivation for our activity in the area of decision-making was not merely extending the standard approaches to deal with fuzzy instead of crisp values, but rather to propose directly a fuzzy-oriented knowledge representation, which is based on the notion of linguistic label [16]–[18]. This idea changes the way of comparing and classifying objects in a fuzzy information system. It is more convenient and computationally not demanding to simply discover classes of objects that share a common characteristic by having the same dominant linguistic values of attributes. We do not apply a standard fuzzy similarity relation for making a detailed analysis of similarity between particular objects of a fuzzy universe. This way we also avoid getting obscured results that depend on the used forms of fuzzy connectives and may be difficult to interpret.

In this paper, we introduce several new concepts into the label-based approach to multi-criteria decision-making. Firstly, we express the preferences of a decision-maker in the form of imaginary alternatives that constitute a reference for comparison and evaluation of real alternatives. Secondly, we define a notion of approximation of the reference alternatives by the real alternatives. This makes it possible to simplify the evaluation of accordance of the real alternatives with the preferences of the decision-maker.

II. Classification of alternatives using fuzzy linguistic labels

It is necessary to recall the basic notions that are used in the presented approach. The process of decision-making is performed by evaluating the alternatives that are characterized by fuzzy criteria. We denote by [16]:

U – a nonempty set (universe) of alternatives (elements),
A – a finite set of fuzzy criteria (attributes),
$\mathbb{V}_a$ – a set of linguistic values of every criterion $a \in A$,
$\mathbb{V}$ – a set of linguistic values of criteria, $\mathbb{V} = \bigcup_{a \in A} \mathbb{V}_a$,
f – an information function, $f : U \times \mathbb{V} \to [0, 1]$, $f(u, V) \in [0, 1]$, for all $u \in U$, and $V \in \mathbb{V}$.

The corresponding family of linguistic values of a fuzzy criterion $a_i \in A$, where $i = 1, \ldots, n$, is expressed as $\mathbb{A}_i = \{A_{i1}, \ldots, A_{in_i}\}$. The membership degrees of any alternative $u \in U$ in the linguistic values of all fuzzy criteria should be assigned under the following conditions [18]:

$$\begin{aligned} \exists A_{ik} \in \mathbb{A}_i \quad & (\mu_{A_{ik}}(u) \geq 0.5, \\ & \mu_{A_{ik-1}}(u) = 1 - \mu_{A_{ik}}(u) \ \vee \\ & \mu_{A_{ik+1}}(u) = 1 - \mu_{A_{ik}}(u)), \end{aligned} \tag{1}$$

$$\text{power}\left(\mathbb{A}_i(u)\right) = \sum_{k=1}^{n_i} \mu_{A_{ik}}(u) = 1. \tag{2}$$

Because of the requirements (1) and (2), every alternative $u \in U$ must have a dominant linguistic value for each fuzzy criterion. Furthermore, due to the requirement (1), every alternative $u \in U$ can take a nonzero membership degree in at most two neighbouring linguistic values.

Introducing the requirement (1) was inspired by practical applications of fuzzy inference systems. It can be observed in a real-world case that an expert usually assigns a nonzero membership degree in only two neighbouring linguistic values of a criterion. This is in accordance with our intuition. We can describe, e.g., an object to be "high" and "middle" to some degree, but not "low" at the same time. Hence, the parameters of membership functions that are applied in fuzzy control systems should be set in such a way that only two neighbouring membership functions can be activated by any crisp input value [19], [20].

The requirement (2) can be perceived as a generalization of the property that can be observed in crisp information systems that constitute a special case of fuzzy information systems. Every element of a crisp universe can have a nonzero membership, which equal to 1, in only one single value of a selected crisp attribute. All the membership degrees in the remaining values of the crisp attribute must be equal to zero. This property of possessing exactly one attribute value can be also expressed by applying the law of excluded middle, which is a crucial principle of the standard bi-valued logic. In a fuzzy information system, an element of the universe can possess a membership degree, which can be any real number in the interval $[0, 1]$, in more than one attribute value. The process of assigning the degree of membership in fuzzy sets is a fundamental issue in applications of the fuzzy set theory. Some researches do not impose strict constraints in this regard. This can be observed especially in the area of neuro-fuzzy systems where the membership degrees are often tuned freely during a training process. Unfortunately, we lose the correspondence to logic in such an arithmetic-oriented approach. Therefore, a well-defined fuzzy information system should satisfy a requirement that can be seen as a counterpart of the law of excluded middle in fuzzy logic.

The requirement (2) is important in real-world applications, e.g., in fuzzy control [19], because it helps to construct a consistent system of fuzzy decision rules that can be easily interpreted. Another example is the fuzzy flow graph approach introduced in [21]. It that case, the requirement (2) has to be be used if we want to retain the flow conservation equations that describe the flow distribution in a fuzzy flow graph.

Both the requirements (1) and (2) allow to significantly simplify the process of constructing a fuzzy information system. Obviously, the formulae (1) and (2) are fully compatible with the standard bi-valued logic in a special case of crisp information systems.

We also require a certain level of membership degree to the dominant linguistic values, which can further restrict the subset of alternatives taken into consideration. Such (positive) alternatives can be easily discovered by finding the unique membership degrees, for all criteria $a \in A$ that exceed a threshold of similarity, denoted by β, that satisfies the inequality: $0.5 < \beta \leq 1$.

Depending on the value of membership degree, and the threshold β, three kinds of linguistic values [17] can be distinguished. For every alternative $u \in U$, and any attribute $a \in A$, we define the set $\widehat{\mathbb{V}}_a(u) \subseteq \mathbb{V}_a$ of positive linguistic values

$$\widehat{\mathbb{V}}_a(u) = \{V \in \mathbb{V}_a : f(u, V) \geq \beta\}, \tag{3}$$

the set $\overline{\overline{\mathbb{V}}}_a(u) \subseteq \mathbb{V}_a$ of boundary linguistic values

$$\overline{\overline{\mathbb{V}}}_a(u) = \{V \in \mathbb{V}_a : 0.5 \leq f(u, V) < \beta\}, \tag{4}$$

and the set $\breve{\mathbb{V}}_a(u) \subseteq \mathbb{V}_a$ of negative linguistic values

$$\breve{\mathbb{V}}_a(u) = \{V \in \mathbb{V}_a : 0 \leq f(u, V) < 0.5\}. \tag{5}$$

We can identify alternatives $u \in U$ that have nonempty sets $\widehat{\mathbb{V}}_a(u)$ for all criteria $a \in A$, according to (3). Those alternatives are marked by distinct labels, which are combinations of their positive linguistic values of criteria.

The set of linguistic labels $\widehat{\mathbb{L}}(u)$ is expressed as the Cartesian product of the sets of positive linguistic values $\widehat{\mathbb{V}}_a(u)$, for all $a \in A$:

$$\widehat{\mathbb{L}}(u) = \prod_{a \in A} \widehat{\mathbb{V}}_a(u). \tag{6}$$

By inspecting the membership degrees of every element u of the universe U in all linguistic values of criteria, we obtain classes (granules) of similar alternatives that share the same linguistic label.

We denote by U_L the subset of those elements u of the universe U that correspond to a linguistic label $L \in \mathbb{L}$, for all fuzzy attributes $a \in A$:

$$U_L = \{u \in U : L(u) = L\}. \tag{7}$$

The subset U_L is called the set of characteristic elements of the linguistic label L.

The linguistic label $L \in \mathbb{L}$ can be expressed as an ordered tuple of positive linguistic values, for all attributes $a \in A$:

$$L = \left(\hat{V}_{a_1}^L, \ldots, \hat{V}_{a_n}^L\right). \tag{8}$$

In the proposed novel approach presented in this paper, we denote by X the universe of real alternatives having the membership degrees assigned by an expert. The alternatives $x \in X$

will be evaluated and ranked with respect to the subjective preferences provided by a decision-maker. In contrast to our previous work, we propose a different way of specifying the preferences of the decision-maker. This is done by introducing ideal (imaginary) reference alternatives that should reflect the expectations of the decision-maker, who must assign their membership degrees in the preferred linguistic values of each fuzzy criterion. The reference alternatives will be denoted by y, and their universe by Y.

We would usually expect the universe Y of imaginary alternatives to contain only one or a small number of elements, because the decision-maker ought to strictly specify his or her requirements. However, in a real-world case, the criteria should not be treated as totaly independent characteristics of alternatives. Rather, we must assume that the decision-maker considers several (possible) versions of ideal alternatives having different combinations of preferred linguistic values of criteria. Moreover, such imaginary alternatives can be treated as solution variants that are not equally desirable.

III. Approximation of alternatives in the domain of linguistic values

In the first step, we determine the set of linguistic labels $\mathbb{L}^E$ of the real alternatives $x \in X$ described by the expert, and the set $\mathbb{L}^D$ of linguistic labels of the reference alternatives $y \in Y$ provided by the decision-maker.

The preferences of the decision-maker can be satisfied only, if we can find linguistic labels that are shared by the expert and the decision-maker, i.e., $\mathbb{L}^E \cap \mathbb{L}^D \neq \emptyset$. Only the real alternatives belonging to the characteristic sets of the common linguistic labels constitute variants of acceptable solutions. All the other real alternatives are not in accordance with the preferences of the decision-maker. In other words, they do not support linguistic labels of the decision-maker, hence, they will be discarded from further consideration.

In the next step, we need to calculate the degree of accordance of the real alternatives with the corresponding reference alternative. We refer to the idea of approximation of sets, which is a fundamental concept of the rough set theory. The notions of the lower and upper crisp set approximations, proposed by Pawlak [1], were extended by many researchers, who developed various generalizations for the case of fuzzy information systems, e.g., [2]–[6].

In approximation of crisp sets a unique indiscernibility relation is used, whereas fuzzy sets can be approximated with the help of a similarity relation. Contrary to the crisp indiscernibility relation, there is no unique way how a fuzzy similarity relation is defined. However, it could be shown [4] that the fuzzy approximations based on the (residual) R-implicators satisfy the largest number of basic properties of rough sets. This fact inspired us to consider R-implicators as a suitable tool for approximation of alternatives in the domain of linguistic values.

Let us define the accordance relation of a real alternative with a reference alternative, with respect to particular linguistic values of criteria. To this end, we represent any alternative x as the fuzzy set $\widetilde{X}$, and any alternative y as the fuzzy set $\widetilde{Y}$ on the domain of linguistic values of criteria. Now, we are able to determine the covering of the set $\widetilde{Y}$ by the set $\widetilde{X}$. The covering will be expressed as a fuzzy set in the domain of linguistic values of criteria, and denoted by $\mathrm{COV}(\widetilde{X}, \widetilde{Y})$.

For a linguistic value A_{ik} of a criterion a_i, we require that the membership degree in A_{ik} of the real alternative x must be at least equal to the membership degree in A_{ik} of the reference alternative y, as a condition to assign the highest possible covering degree (with respect to A_{ik}) of $\widetilde{Y}$ by $\widetilde{X}$:

$$\mu_{\mathrm{COV}(\widetilde{X},\widetilde{Y})}(A_{ik}) = 1 \iff \mu_{\widetilde{Y}}(A_{ik}) \leq \mu_{\widetilde{X}}(A_{ik}). \tag{9}$$

This requirement can be satisfied, when the covering of fuzzy sets is defined by applying residual implicators, which are generally expressed with the help of a t-norm operator T, for $a, b \in [0, 1]$:

$$\mathrm{I}(a, b) = \sup\{\lambda \in [0, 1] : \ \mathrm{T}(a, \lambda) \leq b\}. \tag{10}$$

The R-implicator of Gaines turned out to be the most suitable in our approach, because it takes into account the ratio between its arguments. It has the following form:

$$\mathrm{I}(a, b) = \begin{cases} 1, & \text{if } \ a \leq b, \\ b/a, & \text{otherwise.} \end{cases} \tag{11}$$

By applying the selected R-implicator, we define the covering set $\mathrm{COV}(\widetilde{X}, \widetilde{Y})$ of a nonempty set $\widetilde{Y}$ by the set $\widetilde{X}$ as follows:

$$\mu_{\mathrm{COV}(\widetilde{X},\widetilde{Y})}(A_{ik}) = \begin{cases} \mathrm{I}(\mu_{\widetilde{Y}}(A_{ik}), \mu_{\widetilde{X}}(A_{ik})), \\ \text{if } \ \mu_{\widetilde{X}}(A_{ik}) > 0, \\ 0, \ \text{ otherwise.} \end{cases} \tag{12}$$

On the other hand, we should take also into account the preferences for neighbouring linguistic values of criteria that are specified by the decision-maker. Even in the case of a full covering with respect to a dominant linguistic value, the decision-maker and the expert can assign a nonzero membership of alternatives in different neighbouring linguistic values of a criterion. Therefore, we apply the membership degrees of the reference alternative in the linguistic values of criteria as weighting factors to precisely determine the accordance between the reference and the real alternative. Summarizing, the accordance degree $acc_{A_{ik}}(x, y)$ of a real alternative $x \in X$ with a reference alternative $y \in Y$, with respect to a linguistic value A_{ik} of a criterion $a_i \in A$, is defined as

$$acc_{A_{ik}}(x, y) = \mu_{A_{ik}}(y) \times \mu_{\mathrm{COV}(\widetilde{X},\widetilde{Y})}(A_{ik}). \tag{13}$$

By adding up the accordance degrees $acc_{A_{ik}}(x, y)$ calculated for all linguistic values of a criterion $a_i \in A$, we get the accordance degree $acc_{a_i}(x, y)$ of the alternative x with the alternative y, with respect to a_i:

$$acc_{a_i}(x, y) = \sum_{k=1}^{n_i} acc_{A_{ik}}(x, y). \tag{14}$$

The approximation $\mathrm{ACC}_x(y)$ of the reference alternative $y \in Y$ by the real alternative $x \in X$ is a fuzzy set in the domain of linguistic values of criteria:

$$\mathrm{ACC}_x(y) = \{\, acc_{A_{11}}(x,y)/A_{11}, \ldots, acc_{A_{ik}}(x,y)/A_{ik}, \ldots, acc_{A_{nn_n}}(x,y)/A_{nn_n}\}. \tag{15}$$

One has to remember that the criteria are not equally important in a typical multi-criteria optimization task. This restriction can be formally expressed with the help of weights $w_1, \ldots, w_n$ that can take values from the interval $[0, 1]$, and must add up to 1: $\sum_{i=1}^{n} w_i = 1$. The values of the weights for particular criteria can be freely changed, depending on the choice of the decision-maker.

Finally, we define the measure of acceptance of a real alternative x as follows:

$$acc(x,y) = \sum_{i=1}^{n} w_i \times acc_{a_i}(x,y). \tag{16}$$

The final ranking of alternatives can obtained basing on the value of the acceptance measure, but it should be performed separately for each reference alternative of the decision-maker.

IV. Example

In order to illustrate the presented approach, let us assume a fuzzy information system that contains 10 real alternatives characterized by three fuzzy criteria a_1, a_2, and a_3. All criteria can take three linguistic values and have the same importance, i.e., the weights w_1, w_2, and w_3 are set to $\frac{1}{3}$. The similarity threshold β is equal to 0.6.

For the reference alternatives y_1, y_2, and y_3 provided by the decision-maker (Table II), we get the linguistic labels L_1, L_2, and L_3, respectively. By inspecting the universe X containing the real alternatives, we can find the same linguistic labels and their corresponding sets of characteristic elements X_{L_1}, X_{L_2}, X_{L_3}, respectively:

$$\begin{aligned} L_1 &= (A_{13}A_{22}A_{32}): \quad X_{L_1} = \{x_3, x_{10}\}, \\ L_2 &= (A_{12}A_{23}A_{32}): \quad X_{L_2} = \{x_2, x_6, x_9\}, \\ L_3 &= (A_{13}A_{23}A_{33}): \quad X_{L_3} = \{x_5, x_8\}. \end{aligned}$$

The alternative x_4 has no linguistic label, because all its membership degrees in the linguistic values of the criterion a_2 are below the similarity threshold β. The alternatives x_1, and x_7 do not support the reference labels of the decision-maker. Hence, the alternatives x_1, x_4, and x_7 will be discarded from the solution space.

To demonstrate the details of determining the approximation of the reference alternatives, we select the fuzzy sets $\widetilde{Y}_1(a_1)$, and $\widetilde{X}_3(a_1)$ that represent the alternatives y_1, and x_3, respectively, in the domain of linguistic values of the criterion a_1:

$$\begin{aligned} \widetilde{Y}_1(a_1) &= \{0.00/A_{11}, 0.30/A_{12}, 0.70/A_{13}\}, \\ \widetilde{X}_3(a_1) &= \{0.00/A_{11}, 0.20/A_{12}, 0.80/A_{13}\}. \end{aligned}$$

By applying the formulae 12, 13, and 14, we get the results given in Table III, including the covering set $\mathrm{COV}(\widetilde{X_3}, \widetilde{Y_1})$, the approximation of the reference alternative y_1 by the real alternative x_3, and the accordance degrees with respect to the criteria a_1, a_2, and a_3.

Tables IV, V, VI contain the accordance degrees of the supporting real alternatives with the reference alternatives y_1, y_2, and y_3, respectively. As we can see, the acceptance degrees of the supporting real alternatives are relatively high. However, those three groups of alternatives should be seen as distinct solution variants with separate rankings, and may have different meaning or importance for the decision-maker.

V. Conclusions

Finding an optimal solution by evaluating and ranking the alternatives that are characterized by subjective criteria can be done by utilizing the notion of fuzzy linguistic label. In this paper, we propose to express the subjective preferences of a decision-maker as ideal reference alternatives. They can be represented by respective linguistic labels, in the same way like the real alternatives that constitute the solutions space. By unifying the description in the multi-criteria decision-making process, we are able to define a notion of approximation in the domain of linguistic values of criteria, and to use it for introducing an accordance relation for comparing both kinds of alternatives. The presented method can be a starting point to develop novel hybrid approaches to solving complex multi-criteria decision-making tasks, with respect to mixed subjective and objective criteria.

References

[1] Z. Pawlak, *Rough Sets: Theoretical Aspects of Reasoning about Data*. Boston Dordrecht London: Kluwer Academic Publishers, 1991.

[2] D. Dubois and H. Prade, "Putting rough sets and fuzzy sets together," in *Intelligent Decision Support: Handbook of Applications and Advances of the Rough Sets Theory*, R. Słowiński, Ed. Boston Dordrecht London: Kluwer Academic Publishers, 1992, pp. 203–232.

[3] S. Greco, B. Matarazzo, and R. Słowiński, "Rough set processing of vague information using fuzzy similarity relations," in *Finite Versus Infinite — Contributions to an Eternal Dilemma*, C. S. Calude and G. Paun, Eds. Berlin Heidelberg: Springer-Verlag, 2000, pp. 149–173.

[4] A. M. Radzikowska and E. E. Kerre, "A comparative study of fuzzy rough sets," *Fuzzy Sets and Systems*, vol. 126, pp. 137–155, 2002.

[5] J. T. Starczewski, *Advanced Concepts in Fuzzy Logic and Systems with Membership Uncertainty*, ser. Studies in Fuzziness and Soft Computing. Berlin Heidelberg: Springer-Verlag, 2013, vol. 284.

[6] L. D'eer and C. Cornelis, "A comprehensive study of fuzzy covering-based rough set models: Definitions, properties and interrelationships," *Fuzzy Sets and Systems*, vol. 336, pp. 1–26, 2018.

[7] O. S. Vaidya and S. Kumar, "Analytic hierarchy process: An overview of applications," *European Journal of Operational Research*, vol. 169, no. 1, pp. 1–29, 2006.

[8] S. Greco, M. Ehrgott, and J. R. Figueira, *Multiple Criteria Decision Analysis: State of the Art Surveys*. New York: Springer-Verlag, 2016.

[9] W. Pedrycz, P. Ekel, and R. Parreiras, *Fuzzy Multicriteria Decision-Making: Models, Methods and Applications*. Chichester: John Wiley & Sons Ltd, 2011.

[10] W. Deni, O. Sudana, and A. Sasmita, "Analysis and implementation fuzzy multi-attribute decision making SAW method for selection of high achieving students in faculty level," *International Journal of Computer Science*, vol. 10, no. 1, pp. 674–680, 2013.

[11] C. Kahraman, S. C. Onar, and B. Oztaysi, "Fuzzy multicriteria decision-making: A literature review," *International Journal of Computational Intelligence Systems*, vol. 8, no. 4, pp. 637–666, 2015.

TABLE I
REAL ALTERNATIVES WITH MEMBERSHIP DEGREES ASSIGNED BY THE EXPERT

	a_1			a_2			a_3		
	A_{11}	A_{12}	A_{13}	A_{21}	A_{22}	A_{23}	A_{31}	A_{32}	A_{33}
x_1	0.00	**0.75**	0.25	0.00	0.20	**0.80**	**0.70**	0.30	0.00
x_2	0.00	**0.65**	0.35	0.00	0.30	**0.70**	0.00	**0.75**	0.25
x_3	0.00	0.20	**0.80**	0.00	**0.80**	0.20	0.00	**0.80**	0.20
x_4	**0.80**	0.20	0.00	**0.50**	**0.50**	0.00	0.00	0.30	**0.70**
x_5	0.00	0.25	**0.75**	0.00	0.35	**0.65**	0.00	0.40	**0.60**
x_6	0.15	**0.85**	0.00	0.00	0.25	**0.75**	0.25	**0.75**	0.00
x_7	**0.55**	0.45	0.00	**0.70**	0.30	0.00	**0.60**	0.40	0.00
x_8	0.00	0.35	**0.65**	0.00	0.25	**0.75**	0.00	0.10	**0.90**
x_9	0.00	**0.80**	0.20	0.00	0.15	**0.85**	0.00	**0.65**	0.35
x_{10}	0.00	0.25	**0.75**	0.30	**0.70**	0.00	0.00	**0.75**	0.25

TABLE II
REFERENCE ALTERNATIVES OF THE DECISION-MAKER

	a_1			a_2			a_3		
	A_{11}	A_{12}	A_{13}	A_{21}	A_{22}	A_{23}	A_{31}	A_{32}	A_{33}
y_1	0.00	0.30	**0.70**	0.00	**0.75**	0.25	0.00	**0.85**	0.15
y_2	0.00	**0.80**	0.20	0.00	0.30	**0.70**	0.00	**0.70**	0.30
y_3	0.00	0.30	**0.70**	0.00	0.25	**0.75**	0.00	0.20	**0.80**

TABLE III
ACCORDANCE OF x_3 WITH y_1

	a_1			a_2			a_3		
	A_{11}	A_{12}	A_{13}	A_{21}	A_{22}	A_{23}	A_{31}	A_{32}	A_{33}
x_3	0.00	0.20	0.80	0.00	0.80	0.20	0.00	0.80	0.20
y_1	0.00	0.30	0.70	0.00	0.75	0.25	0.00	0.85	0.15
$\mathrm{COV}(\widetilde{X_3}, \widetilde{Y_1})$	0.00	0.67	1.00	0.00	1.00	0.80	0.00	0.94	1.00
$\mathrm{ACC}_{x_3}(y)$	0.00	0.20	0.70	0.00	0.75	0.20	0.00	0.80	0.15
$acc_{a_i}(x_3, y_1)$		0.90			0.95			0.95	

TABLE IV
ACCEPTANCE DEGREE OF THE ALTERNATIVES OF X_{L_1} WITH y_1

	a_1	a_2	a_3	$acc(x_1, y_1)$	Position
x_3	0.90	0.95	0.95	0.933	1
x_{10}	0.95	0.70	0.90	0.851	2

TABLE V
ACCEPTANCE DEGREE OF THE ALTERNATIVES OF X_{L_2} WITH y_2

	a_1	a_2	a_3	$acc(x_2, y_2)$	Position
x_2	0.85	1.00	0.95	0.9325	2
x_6	0.80	0.95	0.70	0.8165	3
x_9	1.00	0.85	0.95	0.9340	1

TABLE VI
ACCEPTANCE DEGREE OF THE ALTERNATIVES OF X_{L_3} WITH y_3

	a_1	a_2	a_3	$acc(x_3, y_3)$	Position
x_5	0.95	0.90	0.80	0.884	2
x_8	0.95	1.00	0.90	0.950	1

[12] S.-Y. Chou, Y.-H. Chang, and C.-Y. Shen, "A fuzzy simple additive weighting system under group decision-making for facility location selection with objective/subjective attributes," *European Journal of Operational Research*, vol. 189, pp. 132–145, 2008.

[13] J. Kacprzyk, S. Zadrożny, M. Fedrizzi, and H. Nurmi, "On group decision making, consensus reaching, voting and voting paradoxes under fuzzy preferences and a fuzzy majority: A survey and a granulation perspective," in *Handbook of Granular Computing*, W. Pedrycz, A. Skowron, and V. Kreinovich, Eds. Chichester: John Wiley & Sons, 2008, pp. 906–929.

[14] S.-J. Chuu, "Interactive group decision-making using a fuzzy linguistic approach for evaluating the flexibility in a supply chain," *European Journal of Operational Research*, vol. 213, no. 1, pp. 279–289, 2011.

[15] F. Cabrerizo, W. Pedrycz, I. Perez, S. Alonso, and E. Herrera-Viedma, "Group decision making in linguistic contexts: An information granulation approach," *Procedia Computer Science*, vol. 91, pp. 715–724, 2016, www.sciencedirect.com/science/article/pii/S1877050916312455.

[16] A. Mieszkowicz-Rolka and L. Rolka, "A novel approach to fuzzy rough set-based analysis of information systems," in *Information Systems Architecture and Technology. Knowledge Based Approach to the Design, Control and Decision Support*, ser. Advances in Intelligent Systems and

Computing, Z. Wilimowska *et al.*, Eds., vol. 432. Switzerland: Springer International Publishing, 2016, pp. 173–183.

[17] ——, "Labeled fuzzy rough sets versus fuzzy flow graphs," in *Proceedings of the 8th International Joint Conference on Computational Intelligence – Volume 2: FCTA*, J. J. Merelo *et al.*, Eds. SCITEPRESS Digital Library – Science and Technology Publications, Lda, 2016, pp. 115–120, www.scitepress.org/DigitalLibrary.

[18] ——, "Multi-criteria decision-making with linguistic labels," in *Proceedings of the 17th Conference on Computer Science and Intelligence Systems*, ser. Annals of Computer Science and Information Systems, M. Ganzha, L. Maciaszek, M. Paprzycki, and D. Ślęzak, Eds., vol. 30. IEEE, 2022. doi: 10.15439/2022F218 pp. 263–267. [Online]. Available: http://dx.doi.org/10.15439/2022F218

[19] A. Piegat, *Fuzzy Modeling and Control*, ser. Studies in Fuzziness and Soft Computing. Berlin Heidelberg: Springer-Verlag, 2001.

[20] A. Mieszkowicz-Rolka and L. Rolka, "Flow graph approach for studying fuzzy inference systems," *Procedia Computer Science*, vol. 35, pp. 681–690, 2014, www.sciencedirect.com/science/article/pii/S1877050914011156.

[21] ——, "Flow graphs and decision tables with fuzzy attributes," in *Artificial Intelligence and Soft Computing — ICAISC 2006*, ser. Lecture Notes in Artificial Intelligence, L. Rutkowski *et al.*, Eds., vol. 4029. Berlin Heidelberg: Springer-Verlag, 2006, pp. 268–277.

Position Papers of the 18th Conference on Computer Science and Intelligence Systems pp. 103–110
DOI: 10.15439/2023F1990
ISSN 2300-5963 ACSIS, Vol. 36

Comparative Analysis of Low-Code Computation Systems

Anna Rosłan
Warsaw University of Technology
Warsaw, Poland

Michał Śmiałek
Warsaw University of Technology
Warsaw, Poland
0000-0001-6170-443X

***Abstract*—The paper aims to systematically compare computation platforms where the development of custom computation applications is done visually. By this, we mean platforms equipped with a visual language to define the flow of actions or data, thus allowing us to treat them as low-code systems. The chosen platforms include two mature systems: Orange and Azure Machine Learning Studio, and also a newcomer – BalticLSC. For the purpose of the study, two sample computing tasks were created and executed on the three platforms. Based on this, the platforms were compared with each other taking into account the following characteristics: versatility, scalability, user entry barrier, cost of use, availability of documentation, maintainability and extensibility, availability, security, user interface friendliness, and variety of interfaces for input data.**

***Index Terms*—BalticLSC, cloud computing, Orange, Azure, Data Mining, low-code**

I. INTRODUCTION

THE NEED to use low-code platforms comes from the need for companies and enterprises to develop digital tools easily, cheaply and quickly [13]. Currently, digital tools are developed by teams of developers. Team members know programming, programming languages and algorithmics. The development team creates the tool based on the information provided by people who know the field of business requirements and the specifics of the domain in which the tool will be used. Growing demand for the creation and development of digital solutions is creating the need for a large number of programming specialists. Still, there needs to be more of these specialists to meet the growing demand fully. To counteract these shortages and meet the increasing demand for IT solutions, emerging low-code systems aim to develop software in a simple way that allows less skilled workers (in terms of programming skills) to participate in software development tasks related to software development. The paper indicates that the number of publications on low-code systems has increased in recent years, proving the interest in low-code platforms in the commercial market and among the scientific and academic community. The scientific articles surveyed demonstrate that the field of low-code is still being developed and researched and will continue to gain importance.

One such solution is the BalticLSC platform [16], which is a result of a research project to apply the low-code approach to large-scale computations. The claim of BalticLSC is that it should be simple to use and easy to understand by scientists, researchers and other people not proficient in programming. In this paper, we aim at verifying this claim by comparing BalticLSC with two mature platforms that use similar means for defining computation apps – the Orange platform and the Azure Machine Learning Studio. These platforms provide the functionality to create customized computing, data flow and data analysis applications. The platforms also have in common the provision of graphical editors. Additionally, access to these platforms for private use was easy, unlimited and free (except for Azure).

For this purpose, we use a software quality model to compare the features of these three platforms.

II. RELATED WORK

The topic of software quality research has been studied many times [17]. There are many software quality models like ISO 9126, McCall's model, Boehm's model, etc. Most models consider the internal perspective taking into account the software development process. There are also approaches which emphasize the external perspective, i.e. user satisfaction [4], [7], [17]. The proposed quality models consider functional and non-functional software requirements and focus on the user perspective and expectations. Examples of software quality metrics from the user perspective include functionality, usability, reliability, performance, security or support.

Research papers have used different approaches to comparative analysis of applications. Most of them focus on the technical parameters of the systems and the features provided [5], [6], [12]. The analyzed features include the service delivery model (IaaS, PaaS, SaaS), architecture, type of input supported, network configuration, and security-related issues. Another set of benchmarks can include various specific functions provided by the compared tools [8], [14]. This might include, for example, the types of machine learning models available, Decision Trees, neural networks, the ability to load input data in different formats, and types of data visualization (histograms, graphs). Still, such approaches to application analysis provide information about the functions and features of the systems, but provide no data allowing for explicit comparison between tools.

Another comparative application analysis approach is based on comparing features related to user experience and ease of management with each other. Maiya et al. [11] have selected system features such as the learning time required to complete

©2023, PTI

Thematic track: Practical Aspects of and Solutions for Software Engineering

a task, the number of steps taken, ease of use, and the availability of documentation. Each feature was assigned a qualitative measure (e.g., easy, complex) or quantitative measure (10 steps of execution, 10 min). These measures were normalized to a scale of 1-5. This approach to benchmarking allows for a comparative evaluation of systems. Another approach [10] focuses on comparing the computing power of performance platforms and the price of executing an assumed computational problem on selected platforms.

The comparative analysis can also involve solving a selected problem (like heart attack recognition) [9], [15] using the same kind (e.g. machine learning) model on different platforms. Measures of the quality of the services offered are then determined. In the case of using machine learning models, this is, for example, precision and sensitivity. This type of analysis provides information regarding which tool to choose to solve a particular problem (e.g. medical image classification).

III. BalticLSC platform characteristics

BalticLSC [16] is a platform for developing or using large-scale computing applications. The system is designed to provide access to large-scale computing resources to small businesses and institutions that often do not have the resources to buy and maintain the desired infrastructure, as well as the ability to make unused computing resources available to companies and institutions in exchange for financial benefits. The idea is that the system should be easy to use, affordable, and efficient. The user does not need to have specialized knowledge, as by using a graphical language, the users, through the user interface, can design their own applications. The system offers a platform for using ready-made algorithms and applications and creating and sharing one's own designs.

Currently, the system is made available to users in a demo version. This means that some of the functionality has yet to be made available to users. In this paper, the platform will be described in its full version taking into account the functions not available to users to explore the system's full potential. Also, computing centres have not been made available in the demo version. There is one centre located at the Warsaw University of Technology, so the research does not take into account performance tests and speed of execution of test computing programs. The platform was developed as part of a research project co-financed by the European Regional Development Fund. One of the authors of the paper, Michał Śmiałek, is one of the developers of BalticLSC platform, but the study was conducted by someone unfamiliar with the system.

IV. Azure ML Studio characteristics

Azure Machine Learning Studio [2] is a tool that enables one to create, train and deploy machine learning models in the Microsoft Azure cloud. It is a fully managed service that enables machine learning across multiple platforms, including R and Python. Azure Machine Learning Studio allows users to create and deploy machine learning models without programming or machine learning knowledge. Users can import data, perform data mining, produce models, and share results through the user interface. Azure Machine Learning Studio offers many ready-made machine-learning algorithms and allows users to create their own models. Users can also use existing models and customize them to suit their needs.

V. Orange characteristics

Orange [3] is an open-source data mining and business analytics platform that allows users to create and visualise machine learning, neural network, regression, and classification models. It is a drag-and-drop tool, which means users can easily create and modify machine learning models by dragging and dropping feature blocks. The system is provided as a desktop application. The application uses the computing power of the device on which the application is used. Orange includes many built-in machine-learning algorithms and tools for data visualization and presentations of descriptive statistics. The platform also offers many add-ons and plug-ins that allow users to customize the tool to fit their needs.

VI. Methodology for comparative analysis

In order to perform a comparative analysis, two problems were formulated and then solved by the author of this paper on the selected platforms. Following this, the solutions to the problems were compared, taking into account the following features of the systems:

- **Universality**: the openness of the platform to define custom flows and applications.
- **Scalability**: the ability to perform selected tasks on augmented data.
- **User entry barrier**: the number of technologies and tools a user needs to know to use a given platform.
- **Cost of use**: the potential costs associated with performing tasks on the platform.
- **Documentation availability**: the amount and quality of available documentation and information on how to use a given platform.
- **Ease of maintenance and extensibility**: the features provided by the system that enable and facilitate changes, bug fixes, extensions, and enhancements to the application.
- **Availability**: trouble-free operation of the service,
- **Security**: the quantity and quality of the data protection mechanisms used.
- **User interface friendliness**: the ease and clarity of the user interface. The number of steps a user must take to perform a given task.
- **Variety of interfaces for input data**: the ability to provide input data of different formats.

The selection of features for benchmarking was modeled on the software quality models presented in Section II The features selected for comparative analysis reflect an external perspective, i.e. user satisfaction. Another important aspect was to select quality features appropriate to the chosen domain of systems, i.e. low-code platforms.

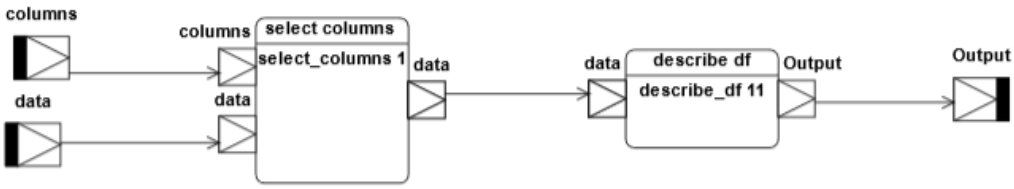


Figure 1. Solution of the first problem in the BalticLSC system

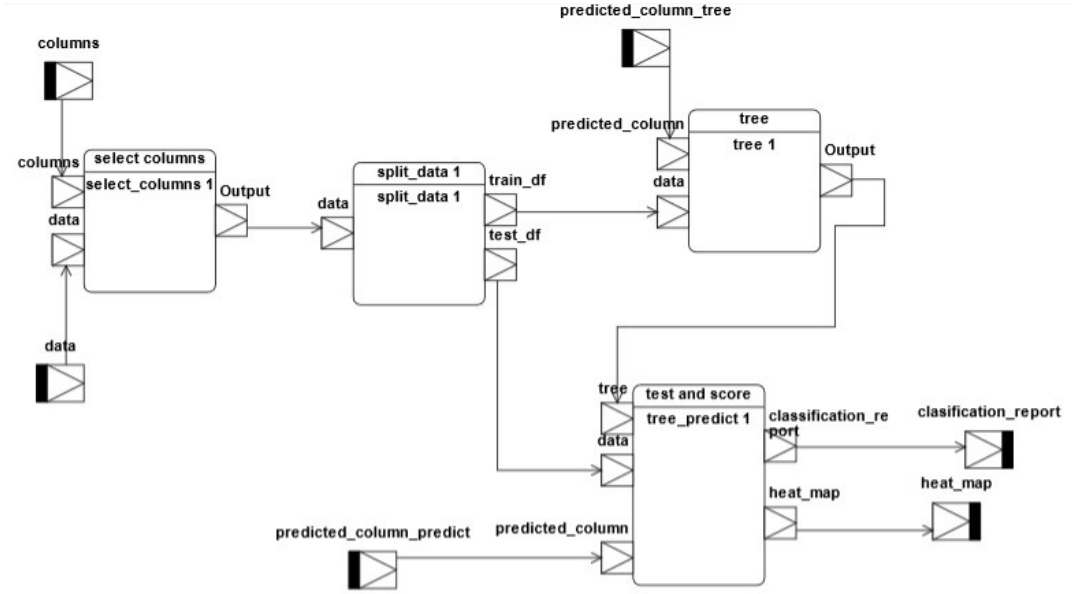


Figure 2. Solution of the second problem in the BalticLSC system (part one)

The first problem to be solved is static analysis of the provided data. We need to create a simple script in Python or another available language that analyzes the provided data. The result of the program is a CSV file that contains the descriptive statistics of the provided data. The second task is to create a decision tree on input data. We divide the data into a test set and a learning set, then create a decision tree and test the model's performance on the test data. The results of the program are files showing the effect of the model, e.g. tree graphics, and analysis of the model prediction, e.g. confusion matrix, and measures of prediction accuracy.

The first task is to check the selected platforms whether they enable the execution of user-defined actions, which checks platform flexibility and versatility. The creation of descriptive statistics action is used as an example of an action that the user must prepare himself. The execution of the task is designed to test the platform's ability to define custom flows freely

Task two is designed to test the execution of an example workflow related to machine learning. Task two contains steps that are typical of an area related to machine learning, i.e. data processing, data partitioning, training the model, and testing the model. Evaluation of the system consists of a comparative analysis of selected platforms for the presented set of features. As a result of the comparative analysis, systems are ranked by awarding first, second and third place.

VII. Implementation of low-code applications

A. *BalticLSC*

The BalticLSC system offers integration with technologies that store data, but it does not have data storage facilities. It was decided to prepare the input data as a CSV file available via an FTP server. The computation results would also be stored on this server. In addition, BalticLSC has no functionality for configuring computation modules through a GUI. Instead, configuration parameters such as what columns to choose for analysis have to be provided as additional input. This allows to select different parameter sets simply by changing the address of the file with the parameters.

Figure 1 shows the solution to the first problem. On the left are graphical representations of the input data, called the data pins. The user can configure input data on the platform by specifying the address (e.g. a URL and a file name) through which the platform obtains the data. In our case, data was provided through an FTP server. The data pin called "data"

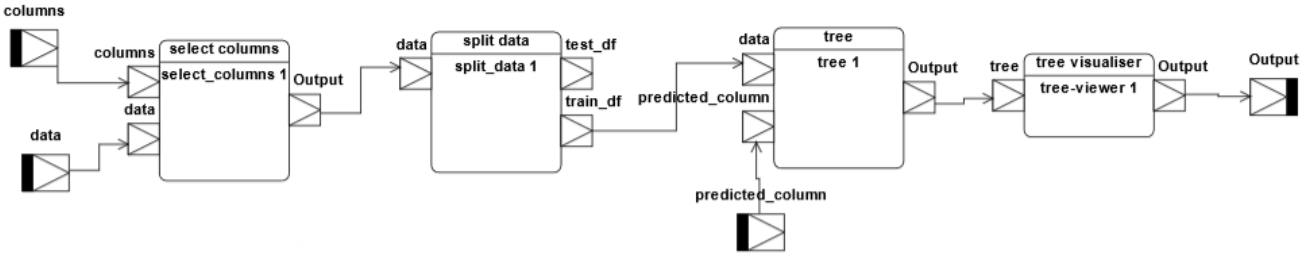


Figure 3. Solution of the second problem in the BalticLSC system (part two)

represents a CSV file with the data to be analysed. The data pin called “columns” represents the list of columns to be analyzed. These two data elements are input to the first module, called “select_columns”. It transforms the input dataset using the selected columns and returns it as the module’s output. The output effect of the module is passed to the next computation module, named ”describe_df”. This module creates descriptive statistics for the transformed data organised into columns. The effect of these computations is passed to the output data pin (“Output”). As for the input data pins, this pin is configured so that the data can be saved to a specific address, which in this case is on an FTP server.

While designing the solution to the second problem, we have noticed that it was not possible to use data coming out of a component iteratively, so two applications had to be created. The first one (Figure 2) accepts the same kinds of data as for the first problem (“data” and “columns”). They are used to create the calculation model. The application has two additional input data pins - “predicted_column_tree” and “predicted_column_predict”. They describe what column from the main input data set (“data”) will be predicted in the machine model. We need two input data pins because a data pin cannot be used more than once. Execution of the application starts with selecting columns for further processing using the “select_columns” module. This results in selecting columns from the main data set to be used for further processing. The result is passed to the “split_data” module, which randomly splits the data into a training and a test set which are output to the “train_df” and “test_df” pins. The training set is passed along with the information about column predictions to the “tree” module. This module creates a decision tree, where the result is a plotted model. This model is passed to the final “test and score” module. Apart from the decision tree, this module receives two inputs – the test dataset “test_df” and information about the predicted column “predicted_column_predict”. The module tests the created model and returns a report (“classification_report”) and a confusion matrix as a heat map (“heat_map”). These two outputs from this application are then stored in the same way as in the first application.

The second application is shown in Figure 3 and is prepared similarly to that from Figure 2. The main difference is the last module – “tree visualizer”. It receives a trained decision tree model as input and creates its graphical representation. The resulting visualisation (diagram) is passed to the output and is stored on an FTP server, as in the other applications.

B. Orange

As in BalticLSC, the execution of tasks in Orange is done within the environment. However, computations are done locally, so there is no need for external storage. In our case, CSV files containing input data were imported directly into the system. The resulting data was accessible directly through a graphical interface.

The application solving the first problem in the Orange notation is shown in Figure 4. Its execution begins with loading the CSV file with data into the “CSV File Import” module. Then, the data is passed to the “Select Columns” module, which selects data contained in selected columns. Unlike for BalticLSC, the columns are not input as a separate file but are defined as direct parameters of the “Select Columns” module. This is done through a GUI and needs to be changed for different computations. After selecting columns, the application runs the “Python script” module to execute a dedicated Python script that creates descriptive statistics for the provided data. The script is defined within the module details. The result of these computations is a data set that is saved to the local machine using the “Save Data” module. The location of the file is given in the details of the module. In addition, it is possible to display data directly in the application using the module “Data Table”. It can be done by showing the module details.

To solve the second problem (see Figure 5), we use the “CSV File Import” and “Selected Columns” modules described above. Data from selected columns is passed to the “Data Sampler” module, which randomly separates the data into a test set and a training set. The training set is passed to the “Tree” module, which creates a decision tree. The trained model is then passed to the “Tree Viewer” module, which creates a graphical representation of the tree. The “Test and Score” module uses the test data to test the tree and returns test information. This information is used by the “Confusion Matrix” module, which creates a confusion matrix as its name suggests. Data passed and produced by the “Tree Viewer”,

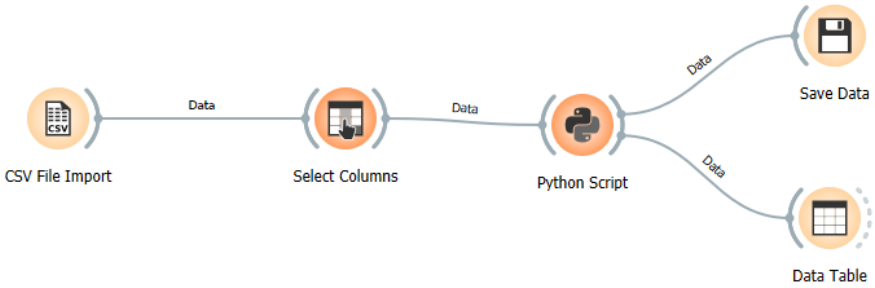


Figure 4. Solution of the first problem in the Orange system

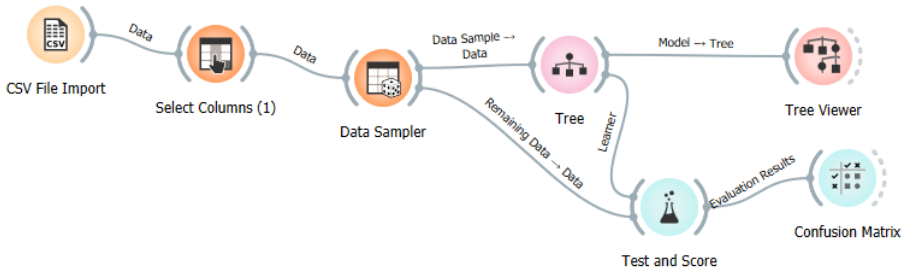


Figure 5. Solution of the second problem in the Orange system

"Confusion Matrix", and "Test and Score" modules can be viewed in the module details. There, one can also specify additional parameters, for example, the predicted column in the "Tree" module.

C. Azure ML Studio

The Azure ML Studio platform operates in the cloud, similar to BalticLSC. We need to create and configure an appropriate service and then design and execute a dedicated solution application. Input data was imported through a GUI into the Azure Blob Storage service, integrated with the platform. As previously, data was contained in CSV files. The output data was visible through the platform's GUI. It was also saved in the Azure data storage services.

The solution to the first problem (Figure 6) contains the input data "household_production" module connected with the "Select columns in Dataset" module. As in the previous cases, this module creates a dataset for selected columns. As in Orange, the list of columns is defined in the details of the module. The resulting data is passed to the "Python Script" module, which executes a script that creates descriptive statistics for the provided data. The script is defined using a graphical interface for editing the module details. The created statistics are passed to the "Export Data" module, which saves the data in a database integrated with the service.

The solution to the second task (Figure 7) starts with the same two modules as in Figure 6. The result is passed to the "Split Data" module, which randomly splits the data into training and test sets. The "Two-Class Boosted Decision Tree" module contains an untrained model, which, together with the training data, is passed to the "Train Model" module. This module trains the input machine learning model on the provided data. The "Score Model" module uses the trained model and the test data to test the tree's predictions. The "Evaluate Model" module calculates various metrics, e.g. precision sensitivity and confusion matrix for the tested model. Finally, these results are stored in a database integrated into the platform.

VIII. Comparative Evaluation of the Systems

Universality. In Azure ML Studio, users only can use the platform's components. The Orange system allows to use the platform's components or create own component, and allows to install additional packages provided by the developers. Both of these systems provide machine learning and artificial intelligence functions. The BalticLSC system can create custom components and applications from any domain and allows to publish the created components. In addition, extensions can be created in any programming language. For this reason, the BalticLSC platform was identified as the most versatile among the respondents. The Azure ML Studio platform fared the worst.

Scalability. The Orange system uses the machine's computing power on which the program is run. The machine's computing power should match the computing power needed to perform the task. Executing a computationally complex task may require adjusting the physical infrastructure so that the available computing power matches the required power needed

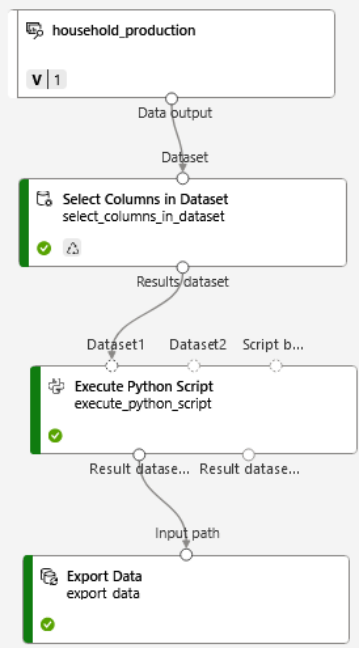


Figure 6. Solution of task one on the Azure ML Studio

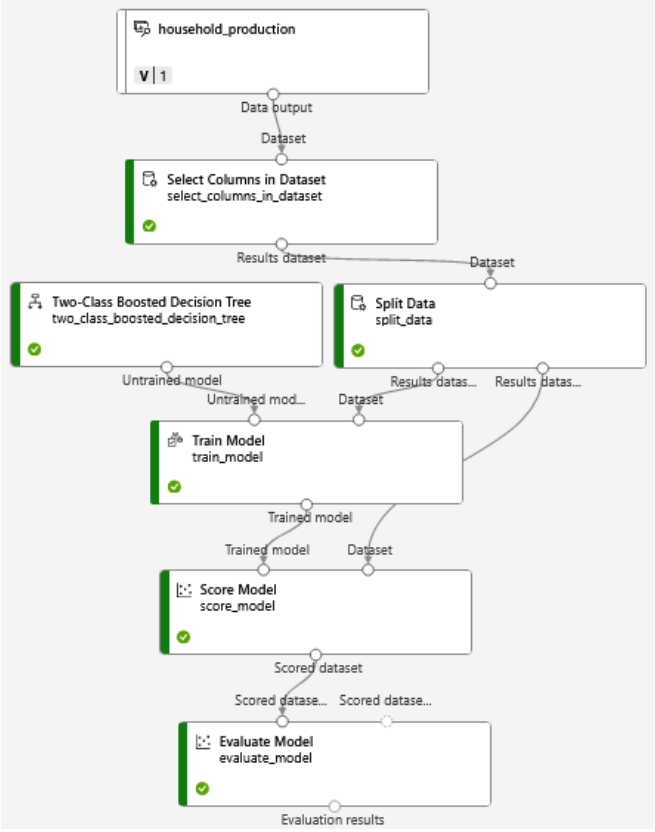


Figure 7. Solution of task two on the Azure ML Studio

to perform the task. The BalticLSC system and Azure ML Studio offer virtual use of computing power through cloud computing. The cloud is easier and generates less cost, so it is tidier. In BalticLSC, choosing the cluster on which to perform the commissioned task is possible. It is also possible to define a range of parameters related to the use of resources, such as GPU, CPU and memory. Since the BalticLSC system is in the demo version, only one cluster is available, and testing the above-described functionalities is impossible. Moreover, BalticLSC offers functions related to the parallelization of computing operations. It is possible to create applications to perform some tasks in parallel.

Azure ML studio also provides a choice of a cluster for processing the task. The computing power is automatically adjusted to the task and can be increased if necessary. The user defines the necessary parameters, such as the maximum/minimum number of clusters and location size and the scaling process itself is performed automatically. The user does not influence the process of task scaling and job scaling, and he only defines specific parameters and limits. Azure ML Studio offers the most features related to the ability to perform compute-intensive tasks. The Orange platform performed the worst.

User entry barrier. To perform tasks on selected platforms, users should have a basic knowledge of machine learning models. In addition, the user should know the Python programming language to create a custom script. To perform tasks in the Orange system, the user must be familiar with the system's components. All interaction with the components, their use and parameterization are done through a graphical interface. In order to create a script in Python, one must become familiar with the "Orange Data Mining Library". In the system, it is possible to import data from a database; thus, basic knowledge of database systems is needed. The amount of time spent learning required to complete the task is the least among the selected applications, so this platform was chosen as the best.

Using the BalticLSC Platform requires knowledge of the components offered by the system. To run the applications, users can configure a computing cluster requiring basic physical infrastructure knowledge. To import data into the application, it is necessary to create a dataset and define the type of access to it. Depending on the chosen technology, basic knowledge of database systems and FTP servers, Azure-DataLake platform, Amazon Simple Storage Service platform is needed. To create a component for BalticLSC (Python script), you need to make it using the Docker tool. For this, users need knowledge of this tool and basic knowledge of containerization. In addition, knowledge of the BalticLSC library is needed to create the code. The variety of technologies and tools for using the system is the greatest. Most of the time was spent learning about the platform and additional tools and technologies, so this platform was rated the lowest.

In order to use the Azure ML Studio platform, one should know about the services offered by the Azure platform. Azure ML Studio requires creating, configuring, and maintaining several services available on the platform. Each of these services

is non-trivial, requiring many parameters to be defined. Azure ML Studio offers a lot of features and capabilities. In order to use this tool, users should have spent the most time learning about the platform, but there was no need to be familiar with other tools and technologies. The learning time for the new tools was estimated based on the time difference between the time it took to complete the entire task and the time it took to create and run the diagram.

Cost of use. Orange ML Studio is a free platform. The BalticLSC system in the demo version does not charge fees, but the full version will require them. There is a fee for using the Azure platform. You pay for using computing power, i.e., running the designed applications. The exact costs depend on the configuration of the computing cluster. Considering this, the Orange platform is the best in terms of cost of use, and the worst is Azure ML Studio.

Documentation availability. The BalticLSC platform has the least extensive documentation. The user can obtain information from the website https://www.balticlsc.eu. The documentation is in pdf form and is 19 pages long. On the YouTube platform, three videos discuss the functions of the system and a video with an example of how to use the platform. The documentation describes the system's architecture, available features and the interface for developing applications. In terms of availability of documentation and community, this platform is rated the worst.

The Orange platform offers an overview of the system through a website at https://orangedatamining.com/. In addition, there are 63 videos provided on the YouTube platform describing the components or showing an example use of the platform. On the website, in addition to a description of features, there are examples of how the system is used.

The Azure platform provides information about Azure ML studio through the website https://learn.microsoft.com/en-us/azure/machine-learning/. The documentation includes a description of the platform's features as well as instructions for creating your environment. It is possible to download extensive documentation in pdf form (2962 pages). In addition, there are many videos about the system on the YouTube platform created by the creators as well as others. Access to the documentation has been rated as the best for Azure ML Studio.

Ease of maintenance and extensibility. All of the platforms studied provide application expansion through a graphical interface. Adding more components is done by selecting a component and dragging it with the mouse to the application development view. The Orange system has no features to support making changes, extending and enhancing applications, so this system was rated the worst. In the BalticLSC system, components and applications are versioned. Operating on versions makes the process of maintaining applications easier. The Azure ML studio platform provides features that double down on sharing components and pipelines with other users, scheduling and automatically running pipelines. In addition, Azure offers features related to model monitoring and automatic model training. This system was rated the best in terms of ease of project maintenance.

Availability. The Azure platform, which includes the Azure ML studio under study, provides 99.99% availability to Virtual Machines. Azure implements many practices including inter-center data redundancy, backup. This platform is rated the best in terms of service availability.

The BalticLSC platform provides the possibility to select a computing cluster, which makes it possible to operate on the remaining clusters in case of failure of one cluster. The system does not store user data on its own. Data is stored in external services, which means there is no risk of data loss in case of system failure.

The Orange platform is entirely dependent on the efficiency and availability of the physical infrastructure on which it is run, which is why it is rated as the worst in terms of availability.

Security. The Orange system has no security-related mechanisms implemented. The user should secure the physical infrastructure on which the program is run. Therefore, this system is rated as the worst in terms of security.

The BalticLSC system is designed in such a way that it does not store user-supplied data or output data from the operation of the program. In addition, the use of the Docker tool allows full separation of the executed program from other elements of the system. The system uses an encrypted https protocol. The system does not provide data on how the computing clusters are protected from unauthorized access or attacks.

ML Studio provides the ability to define and manage user permissions, which increases security. Azure uses an encrypted https protocol. Application developers follow many practices [1] to ensure the security of the physical infrastructure against unauthorized access or attacks. The Azure system has been rated as the best in terms of security.

User interface friendliness. The Orange platform provides the most straightforward user interface. The user can perform a task on a single view of the application entirely through the GUI. The least number of steps are taken to perform the designed tasks.

A more complex system is the BalticLSC system. To perform a task, the user has to use four separate views to define input data, select (or design) applications, launch applications, and track the applications' progress.

The most complex system in terms of the user interface is the Azure ML Studio platform. Setting up the platform requires creating several services and configuring platform parameters. Performing the task requires using four views of the application, providing input data, creating an experiment, designing the application, running the application. The complicated user interface is due to the large number of features offered by the platform as well as the ability to customize the platform.

Variety of interfaces for input data. Orange's system allows for data transfer in three ways:

- by uploading data from the system or network address,
- downloading data from a relational database,
- self-creating data via the application interface.

Table I
PLATFORM COMPARISON SUMMARY

	Azure ML	Orange	BalticLSC
Universality	III	II	I
Scalability	I	III	II
User entry barrier	II	I	III
Cost of use	III	I	II
Documentation availability	I	II	III
Ease of maintenance	I	III	II
Availability	I	III	II
Security	I	III	II
User friendly interface	III	I	II
variety of interfaces for input data	II	III	I

BalticLSC allows data transfer in five different ways:

- by cloud services (Amazon S3, Azure Data Lake),
- FTP server,
- uploading data from the system,
- downloading data from relational databases,
- downloading data from NoSQL databases.

The Azure ML Studio platform allows data transfer in four ways:

- uploading data from the system directly into the application via the GUI or command line.
- downloading data from relational databases,
- via cloud services (Amazon S3, Azure Data Lake).

BalticLSC offers the most variety of data delivery options, and Orange the least.

IX. SUMMARY

Table I summarises the results of our analysis. We have compared two mature computation platforms with a new system stemming from a research project. It should also be noted that Azure ML Studio and Orange are dedicated to specific application domains. Their computation capabilities can be extended with simple Python scripts. When comparing them to BalticLSC, we have, in fact, reduced the capabilities of BalticLSC to handle only such simple Python procedures. However, this system allows executing containers with code of any size and complexity and written in any language.

The comparative analysis allowed us to identify each application's unique features, advantages and disadvantages. Orange stood out for its intuitive interface and ease of use. Azure ML Studio offers advanced customization and a wide range of features to help organize work, which attracts users with more advanced needs. BalticLSC excels in the universality of use and the ability to create and share custom components.

Evaluating software quality is a challenging and complex process influenced by subjective factors like user experience and expectations, personal preferences, etc. Some wants are mainly based on subjective feelings like user interface friendliness, and others are less like the cost of use. Therefore, depending on the sample group, the results of the same analysis may vary.

Having in mind these differences, it can be noted that it is not possible to isolate a platform that would be unequivocally the best. Each of the studied systems has its strengths and weaknesses that contribute to their quality which is a multidimensional concept. The studied platforms have features that are attractive to different groups of users. Azure ML Studio outperforms the other systems in documentation, ease of maintenance, availability and security. However, Orange and BalticLSC dominate in such criteria as cost-effectiveness, learnability and universality.

REFERENCES

[1] Azure facilities, premises, and physical security. https://learn.microsoft.com/en-us/azure/security/fundamentals/physical-security. Accessed: 2023-03-10.

[2] Azure Machine Learning documentation. https://learn.microsoft.com/en-us/azure/machine-learning. Accessed: 2023-03-10.

[3] Orange data mining documentation. https://orangedatamining.com/docs. Accessed: 2023-03-10.

[4] Anas Bassam Al-Badareen, Mohd Hasan Selamat, Marzanah A Jabar, Jamilah Din, Sherzod Turaev, and S Malaysia. Users' perspective of software quality. In *The 10th WSEAS international conference on software engineering, parallel and distributed systems (SEPADS 2011)*, pages 84–89. World Scientific and Engineering Academy and Society (WSEAS) Cambridge, 2011.

[5] Meenakshi Bist, Manoj Wariya, and Amit Agarwal. Comparing delta, open stack and xen cloud platforms: A survey on open source iaas. In *2013 3rd IEEE International Advance Computing Conference (IACC)*, pages 96–100, 2013.

[6] C. Höfer and G. Karagiannis. Cloud computing services: Taxonomy and comparison. *Journal of Internet Services and Applications*, 2:81–94, 01 2010.

[7] Amna Ikram, Isma Masood, Tahira Sarfraz, and Tehmina Amjad. A review on models for software quality enhancement from user's perspective.

[8] A. Jovic, K. Brkic, and N. Bogunovic. An overview of free software tools for general data mining. In *2014 37th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, pages 1112–1117, 2014.

[9] Sarangam Kodati and R Vivekanandam. Analysis of heart disease using in data mining tools orange and weka. *Global journal of computer science and technology*, Feb 2018.

[10] Charlotte Kotas, Thomas Naughton, and Neena Imam. A comparison of amazon web services and microsoft azure cloud platforms for high performance computing. In *2018 IEEE International Conference on Consumer Electronics (ICCE)*, pages 1–4, 2018.

[11] Madhavi Maiya, Sai Dasari, Ravi Yadav, Sandhya Shivaprasad, and Dejan Milojicic. Quantifying manageability of cloud platforms. In *2012 IEEE Fifth International Conference on Cloud Computing*, pages 993–995, 2012.

[12] Junjie Peng, Xuejun Zhang, Zhou Lei, Bofeng Zhang, Wu Zhang, and Qing Li. Comparison of several cloud computing platforms. In *2009 Second International Symposium on Information Science and Engineering*, pages 23–27, 2009.

[13] Niculin Prinz, Christopher Rentrop, and Melanie Huber. Low-code development platforms-a literature review. In *AMCIS*, 2021.

[14] Venkateswarlu Pynam, R Spanadna, and Kolli Srikanth. An extensive study of data analysis tools (Rapid Miner, Weka, R Tool, Knime, Orange). *International Journal of Computer Science and Engineering*, 5:4–11, 09 2018.

[15] Ritu Ratra and Preeti Gulia. Experimental evaluation of open source data mining tools (weka and orange). *International Journal of Engineering Trends and Technology*, 68(8):30–35, 2020.

[16] Radosław Roszczyk, Marek Wdowiak, Michał Śmiałek, Kamil Rybiński, and Krzysztof Marek. Balticlsc: A low-code hpc platform for small and medium research teams. In *2021 IEEE Symposium on Visual Languages and Human-Centric Computing (VL/HCC)*, pages 1–4, 2021.

[17] Jagannath Singh and Nigussu Bitew Kassie. User's perspective of software quality. In *2018 Second International Conference on Electronics, Communication and Aerospace Technology (ICECA)*, pages 1958–1963, 2018.

Position Papers of the 18th Conference on Computer Science and Intelligence Systems pp. 111–117
DOI: 10.15439/2023F2798
ISSN 2300-5963 ACSIS, Vol. 36

Evaluation of selected Cardinality Pattern functions and linguistic variables applied to authors dominant discipline classification

Lukasz Szymula
0000-0001-8714-096X
(1) Faculty of Mathematics and Computer Science
(2) Center for Public Policy Studies, Adam Mickiewicz University in Poznan, Poland
(3) Department of Computer Science, University of Colorado Boulder, United States
Email:
lukasz.szymula@amu.edu.pl

Krzysztof Dyczkowski
0000-0002-2897-3176
Faculty of Mathematics and Computer Science,
Adam Mickiewicz University in Poznan, Poland
Email:
krzysztof.dyczkowski@amu.edu.pl

***Abstract*—The ongoing study aimed to investigate the impact of utilizing intelligent counting algorithms to determine the dominant discipline of authors. This paper addresses the issue of ambiguously assigning disciplines to authors, which has become a prevalent problem. The methodology section outlines the approach employed in this study, including the utilization of intelligent counting, cardinality pattern functions, and evaluation metrics. In the results section, we present the findings of the study, demonstrating that by employing specific Cardinality pattern functions and linguistic variables, we were able to achieve a return that surpassed the number of disciplines unambiguously determined for authors by up to 30%, surpassing the results obtained using well-known methods.**

***Index Terms*—intelligent counting, cardinality pattern functions, science of science, determining disciplines**

I. Introduction

IN THE field of Scientometrics, research is conducted at multiple levels, focusing on various units of analysis such as publications or researchers. These levels encompass diverse groups, including countries, disciplines, gender, age groups, sources, and research metrics. To ensure the selection of appropriate observation sets, it is crucial to assign unambiguous values to each observation. Ongoing research conducted at the Center for Public Policy Studies, Adam Mickiewicz University, indicates the need to exclude certain observations from the study due to the absence of specific values. Additionally, the challenge arises from the inability to assign a single, definitive value to each observation."

While for scientific databases in the case of some attributes there may be problems in indicating a specific value (complete impossibility to determine the value or appearance of outlier observations that underestimate the results), in the case of other dimensions alternative ways of determining the value can be considered. In Scientometrics, there are metrics for ranking publications and authors. They take values such as number of citations, number of publications, journal percentile, journal indices, author indices, and others at different depth level [7]. Scientific metrics allow ranking selected observations for the purpose of conducting further evaluation, and the prestige of the studied entity is determined by their placement. There are many metrics (CiteScore, FWCI, Percentile, H-Index, Collaboration metrics, etc.) and university rankings (Academic Ranking of World Universities, CWTS Leiden Ranking, RUR World University Rankings, etc.) [8], [13], [23], where terms like high, highest, most popular, medium, low, lowest, worst are used. These are imprecise terms, which are considered and modeled by the field of artificial intelligence: fuzzy set theory and linguistic variables proposed by L.A Zadeh [27].

II. Problem specification

For authors discipline determination Abramo, Aksnes, and D'Angelo, who defined the Web of Science subject category for each Italian and Norwegian professor in their sample [1] and Kwiek and Roszka [18] and Boekhout, van der Weijden and Waltman [5] who defined Scopus ASJC discipline for Polish scientists and global scientists, respectively, used an approach where they determined the modal value based on journals ASJC classification code. This method for one of the largest scientific database (Scopus), allowed 24,938,113 of the 36,010,088 (around 69%) authors to be classified into one dominant discipline thus leaving the remaining 11,071,975 (around 31%) with more than one discipline assigned (Tab. 1.). These 31 percent of observations are tend to be excluded from the samples. As an alternative to use of disciplines assigned to scientific journals and modal value assignment,

This work is the result of research project No. 2019/35/0/HS6/02591 funded by the National Science Center Poland and supervised by Professor Marek Kwiek.

©2023, PTI

Topical area: Advanced Artificial Intelligence in Applications

techniques based on machine learning taking into account abstracts and keywords of publications have been used. Daradkeh M, Abualigah L, Atalla S, Mansoor W. have done paper field classification for Scopus, ProQuest, and EBSCOhost datasets using convolutional neural networks with titles, abstracts, keywords for papers and journal titles as a features [9]. Sood, S.K., Kumar, N. & Saini, M. have used the set of 7 thousand papers from Scopus database and clustering techniques based on keywords and VOSViewer environment [21]. Meen C. K., Seojin N., Fei W., Yongjun Z. performed graph analysis and text mining for keywords from 10 thousand articles from Web of Science database [19].

TABLE I.
DISTRIBUTION OF THE NUMBER OF SCIENTISTS BY NUMBER OF ASSIGNED DISCIPLINES

Number of disciplines	Number of authors
1	24,938,113
2	7,655,307
3	2,538,076
4	615,292
5	198,175
6	48,216
7	11,773
8	3,528
9	1,382
10	178
11	36
12	11
13	1
Total	36,010,088

With the occurrence of imprecise terms in the area of Scientometrics, a variety of possibilities arise to implement this algorithm. One way to measure sample set cardinality is to consider intelligent counting (Sigma f-Count) with various cardinality pattern functions proposed by Wygralak M. [26] and Dyczkowski K. [11]. Cases where author have more than one dominant discipline create space to introduce the fuzzy logic conceptual apparatus in order to increase the number of unambiguously identified authors.

The purpose of the ongoing study was to investigate the impact of intelligent counting usage in algorithm to determine authors dominant discipline. The study compares the sets of dominant author disciplines implemented using the crisp set approach and using fuzzy sets approach. The research questions are as follows:

1. How does the use of intelligent counting effect the number of uniquely classified authors?
2. Which linguistic variables, terms and cardinality pattern functions are meaningful in acquiring more unambiguously classified authors?
3. Does the result from using intelligent counting assign the same classes as classical approach?
4. Are there cardinality pattern functions that in any case return less observations than the approach using crisp sets.

III. METHODS

The bibliometric database Scopus from the ICSR Lab platform has been used for the study. Access to the database was granted through a collaboration between AMU's Center for Public Policy Studies and the International Center for the Study of Research (ICSR Lab), Elsevier, established in November 2020. The ICSR Lab allows access to the Scopus bibliometric database via the Databricks platform and retrieves results in aggregated form. Computations in the Databricks platform were based using cluster in standard mode with Databricks Runtime version 11.2 ML, Apache Spark technology v3.3.0, Scala v2.12, and instance i3.2xlarge with 61 GB Memory, 8 Cores, 1-6 workers for worker type and instance c4.2xlarge with 15 GB Memory, 4 Cores for Driver type. The execution time for all scripts took approximately 2 hours. Our sample included a set of authors and their dominant disciplines determined using an approach commonly used in Scientometrics (crisp sets, referred to by us as the base approach) and applying methods known from intelligent counting (selected cardinality pattern functions). The results were then evaluated following selected evaluation metrics.

A. Method for determining author's dominant discipline

The rule for determining author's dominant discipline was based on using a set of publications from the ICSR Lab platform. Each publication had its own unique identifier, a list of authors, a list of disciplines assigned to the journal from which the publication had come, and the variables Citation, FWCI 4y, FWCI 5y, FWCI NoWindow, Team size and Percentile. The Citation variable represented the total number of citations of the publication, the FWCI variables represented citation indices (gained up to 4, 5 years after the release date of the publication or without time limitation) normalized to the scientific discipline. Team size represented the number of authors in the publication, and the Percentile variable measured the percentile value of the CiteScore metric from the journal assigned to the publication. To determine an author's dominant discipline in the base approach for each author, the number of publications for each discipline that author had been counted. In the case of intelligent counting, each record was assigned to an appropriate membership degree based on the discipline, linguistics variable and term, and then in each discipline that the author subserved, the membership degrees were summed by relying on the appropriate cardinality pattern function and Sigma f-Count function. Then, for each author, only those disciplines were selected for which the number of publications (and Sigma *f*-Count score, respectively) were the highest. Multidisciplinary was excluded from the collection due to the fact that it is not a scientific discipline. The result was a set containing author identifier and his discipline. For the presented approach, the number of disciplines for each observation was greater than or equal to one.

B. Cardinality pattern functions and Sigma f-Count

For the purposes of this study to calculate the number of publications for each authors discipline we have used the sigma *f*-Count cardinality of a fuzzy set defined as:

$$\forall A \in FFS : sc_f(A) = \sum_{x \in supp(A)} f\big(A(x)\big),$$

where FFS is Family of all Fuzzy Sets, *f* is a cardinality pattern function, *sc* is scalar cardinality and *A(x)* is interpreted as degree of membership of x to a fuzzy set A [26], [11]. As the cardinality pattern functions we decided to select four functions from the two patterns: counting by thresholding and counting by thresholding and joining.

1. $f_{1,t,p}$, where $t \in [0, 1]$ and $p \geq 0$. Called as counting by thresholding and joining by Wygralak [26]

$$f_{1,t,p}(x) = \begin{cases} x^p, & a \geq t, \\ 0, & otherwise. \end{cases}$$

2. $f_{2,t,p}$, where $t \in [0, 1]$ and $p \geq 0$. Called as counting by thresholding by Wygralak [26]

$$f_{2,t,p}(x) = \begin{cases} 1, & a \geq t, \\ x^p, & otherwise. \end{cases}$$

For cardinality patterns above we decided to use these in two combinations by p and 5 combinations by t which we named $f_{3,t}$ for $f_{1,t,p}$ with $p = 1$; $f_{4,t}$ for $f_{1,t,p}$ with $p = 2$; $f_{5,t}$ for $f_{2,t,p}$ with $p = 1$ and $f_{6,t}$ for $f_{2,t,p}$ with $p = 2$, where $t \in \{0, 0.2, 0.4, 0.6, 0.8\}$ giving 20 functions for each term of each linguistic variable (that is 360 calculations in total; 6 linguistic variables, 3 terms, 5 thresholds, 4 cardinality pattern functions) (Fig. 1.) [26], [11].

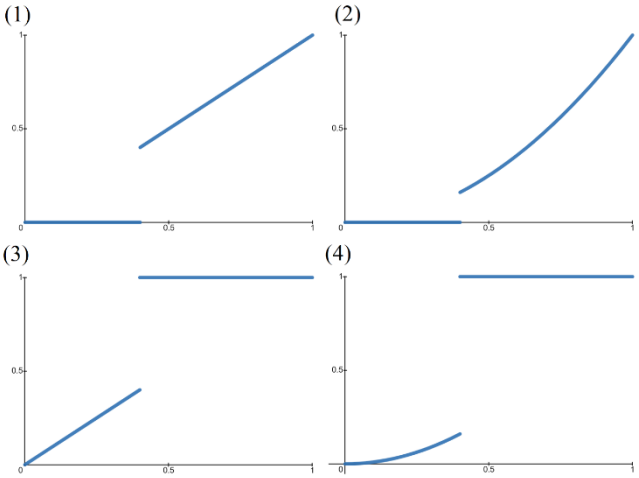


Figure 1. (1) Cardinality pattern $f_{1,t,p}$ for $t = 0.4$, $p = 1$ (as $f_{3,t}$), (2) Cardinality pattern $f_{4,t}$ for $t = 0.4$, $p = 2$ (as $f_{4,t}$), (3) Cardinality pattern f_2 for $t = 0.4$, $p = 1$ (as $f_{5,t}$), (4) Cardinality pattern f_2 for $t = 0.4$, $p = 2$ (as $f_{6,t}$)

C. Membership function modeling

Six linguistic variables (Citation, FWCI 4y, FWCI 5y, FWCI NoWindow, Percentile and Team size) with three terms (Low, Medium, High) in a universe covering the interval [maximum, minimum] for each linguistic variable in the Scopus database have been proposed for the determination of membership functions. The determination of the membership function was based on the same rule for each linguistic variable. In the case of term "high", the membership function was given a value 1 for the top 10 percentile of the variable's value. For the remaining 90 percent, it was an increasing linear function. For term "low", the negation of the "high" membership function was assigned. Term Medium was the minimum of the membership degree values for term "low" and term "high". Due to the patterns/differences that occur in Scientometrics for the given linguistic variables and disciplines, each linguistic variable was modeled for each discipline separately. Two characteristic points were required to establish the membership function. The first point was the minimum for a given term (for Citation, FWCI 4y, FWCI 5y, FWCI NoWindow it was assumed to be 0, for Percentile and Team size it was assumed to be 1). To determine the values of the 90th percentile for linguistic variables, calculations were performed on a set of Scopus publications from the ICSR Lab platform (Tab. 2.).

D. Evaluation metrics

The most popular evaluation metrics used in classification algorithms were used to compare the sets obtained using the base approach and the approach using intelligent counting: Accuracy, Precision, Specificity, F1 and Matthew's correlation coefficient (MCC). The MCC metric was used due to the multi-class nature of classification, which in these cases provides a better measure of quality than Accuracy. The MlLib library for PySpark (MulticlassClassificationEvaluator class) was used to calculate the above metrics. The metrics Accuracy, Precision, Specificity, F1 were available as attributes and due to its multi-class classification nature represent their weighted average score. Due to the limitations of the MlLib library and the large dataset, the determination of MCC was based on the Macro-Averaging method for which the values TP, TN, FP, FN (also obtained by the library) and the equation for binary classification have been used. As the sets that were the subject of comparison, the results from the base approach were always used as the first set, and as the second set, respectively, each subsequent result from the application of each cardinality pattern function. Only subsets in which both the base approach and the intelligent counting approach succeeded in assigning one of the 26 classes (ignoring null values in both the first and second sets) were selected for calculation of evaluation metrics.

A supplemental measure (Set Increase) has been added to determine the percentage of observations received relative to the base approach. This measure accounted for the percentage of difference between the number of unambiguously classified observations using the intelligent counting approach and the number of unambiguously classified observations using the baseline approach to the number of observations obtained using the baseline approach (N=24,938,113). In other words, by how many percent more or less observations were successfully classified by the chosen approach than by using the base approach.

D. Classes (scientific disciplines)

The assignment of authors to scientific disciplines was based on the ASJC (All Science Journal Classification). 26 classes were used for the study. There were 27 disciplines in the ASJC listing, but the Multidisciplinary class was excluded due to the fact that it is not a scientific discipline by itself. The following classes were used for the study: AGRI, agricultural and biological sciences; ARTS, arts and humanities; BIOC, biochemistry, genetics, and molecular biology; BUSI, business, management and accounting; CENG, chemical engineering; CHEM, chemistry; COMP, computer science; DECI, decision sciences; DENT, dentistry; EART, earth and planetary sciences; ECON, economics, econometrics and finance; ENER, energy; ENGI, engineering; ENVIR, environmental science; HEAL, health professions; IMMU, immunology and microbiology; MATE, materials science; MATH, mathematics; MEDI medicine, NEURO, neuroscience; NURS, nursing; PHARM, pharmacology, toxicology, and pharmaceutics; and PHYS, physics and astronomy, PSYC, psychology; SOCI, social sciences; VETE, veterinary.

IV. Results

In this section result has been discussed only for these Cardinality pattern functions where the number of the unambiguously assigned authors was bigger than using base approach. For Card. Pattern function $f_{5,t}$ no increase has been noted for every linguistic variable. The full results with negative Set Increase is presented on GitHub:
(link: https://github.com/lukaszszy/fedcsis-evaluation-cardinality-pattern-functions-authors-dominant-discipline-classification).

For the linguistic variable "Citation," a set increase was noted for all three terms, with the most for the “low” term. For f_3 and f_4, set increase was seen for each threshold. For f_6 only for limits 0.4 to 0.8. For term “mid” and “high” for f_3 and f_4, an increased number of observations was noted only when there was no threshold application. A significant difference can be observed between Accuracy and MCC for the “low” and “high” term. In the case of favoritism for highly cited papers, on average, more than 50 percent of researchers have classified another discipline ($f_{4,0.0}$, $f_{6,0.6}$ and $f_{6,0.8}$); giving a set increase of about 12-13 percent. The term “low” received the largest set increase of more than 30 percent, and results above 20 percent were obtained by 8 of the 13 cardinality pattern functions (Fig. 2). This is explained by the fact that there are more researchers and publications with low citations in the Scopus database than researchers and publications with a high number of citations (on the skewed distribution of publications and citations, see Albarrán et al. [2]; Carrasco and Ruiz-Castillo [6]; Ruiz-Castillo and Costas [20]).

For the FWCI metric, the number of terms for which positive set increases were achieved was incremental. For the "FWCI 4y" variable, positive results were achieved only for the low term, for "FWCI 5y" positive results were achieved for low and “mid”. For "FWCI NoWindow," positive results were achieved for all three terms. For all three FWCI variants, Accuracy and MCC results showed that classification was more similar to base classified disciplines than for the Citation variable (Dominant values above 0.8 and several Cardinality

Term	Card. Pattern f	Accuracy	Precision	Recall	Specificity	F1	MCC	Set increase
low	f3, 0	0.916	0.921	0.916	0.998	0.918	0.916	30.91
	f3, 0.2	0.923	0.928	0.923	0.997	0.925	0.923	29.38
	f3, 0.4	0.902	0.909	0.902	0.997	0.905	0.903	26.74
	f3, 0.6	0.853	0.863	0.853	0.995	0.857	0.857	21.60
	f3, 0.8	0.742	0.758	0.742	0.989	0.749	0.754	9.02
	f4, 0	0.889	0.897	0.889	0.996	0.892	0.891	30.91
	f4, 0.2	0.889	0.896	0.889	0.996	0.892	0.891	29.37
	f4, 0.4	0.887	0.895	0.887	0.996	0.890	0.888	26.73
	f4, 0.6	0.852	0.862	0.852	0.995	0.856	0.855	21.60
	f4, 0.8	0.742	0.758	0.742	0.989	0.749	0.754	9.02
	f6, 0.4	0.933	0.937	0.933	0.998	0.935	0.933	1.84
	f6, 0.6	0.915	0.921	0.915	0.997	0.918	0.916	5.50
	f6, 0.8	0.897	0.904	0.897	0.997	0.900	0.898	12.28
mid	f3, 0	0.823	0.828	0.823	0.988	0.825	0.822	11.24
	f4, 0	0.642	0.654	0.642	0.971	0.648	0.650	11.28
	f6, 0.2	0.810	0.816	0.810	0.987	0.813	0.811	2.03
	f6, 0.4	0.607	0.622	0.607	0.969	0.614	0.618	9.45
	f6, 0.6	0.642	0.654	0.642	0.971	0.648	0.650	11.28
	f6, 0.8	0.642	0.654	0.642	0.971	0.648	0.650	11.28
high	f3, 0	0.671	0.667	0.671	0.973	0.669	0.675	13.73
	f4, 0	0.498	0.485	0.498	0.962	0.491	0.519	13.72
	f6, 0.2	0.683	0.685	0.683	0.975	0.684	0.689	3.06
	f6, 0.4	0.633	0.631	0.633	0.970	0.631	0.641	8.95
	f6, 0.6	0.500	0.488	0.500	0.963	0.494	0.522	11.50
	f6, 0.8	0.499	0.486	0.499	0.962	0.492	0.520	12.89

Evaluation metrics 0.485 0.998 Set increase [%] 1.84 30.91

Figure 2. The results of the evaluation metrics of the linguistic variable “Citation” for the Cardinality pattern functions, where the number of uniquely identified observations exceeded the number of observations from the base approach.

Pattern functions with result above 0.7). The largest set increase (25.22%) was recorded for FWCI 4y AND function f3,0.0. Slightly more observations (26.44%) were noted for the choice of FWCI 5y variable. For FWCI NoWindow, the number of observations over the number of observations in the baseline approach was the highest (30.30% for f4,0.0) (Fig 3., Fig 4. Fig 5.). The prerequisite for obtaining a greater variation in Sigma f-Count values for the 3 time restrictions for the FWCI indicator is the possibility of gathering a larger citation summation; as an additional 1 year for FWCI 5y compared to FWCI 4y and unlimited time for gathering citations in the case of no restriction for years (see Baas et al. [3]; de Moya-Anegón et al. [10]).

Term	Card. Pattern f	Accuracy	Precision	Recall	Specificity	F1	MCC	Set increase
low	f3, 0	0.917	0.922	0.917	0.997	0.919	0.917	25.22
	f3, 0.2	0.906	0.912	0.906	0.997	0.908	0.906	22.92
	f3, 0.4	0.845	0.854	0.845	0.995	0.849	0.850	18.89
	f3, 0.6	0.840	0.853	0.840	0.994	0.846	0.844	11.27
	f4, 0	0.899	0.907	0.899	0.996	0.903	0.900	25.19
	f4, 0.2	0.899	0.906	0.899	0.996	0.902	0.900	22.89
	f4, 0.4	0.842	0.851	0.842	0.995	0.846	0.847	18.86
	f4, 0.6	0.847	0.859	0.847	0.994	0.853	0.851	11.26
	f6, 0.4	0.926	0.929	0.926	0.998	0.927	0.926	1.85
	f6, 0.6	0.916	0.921	0.916	0.997	0.918	0.916	6.71
	f6, 0.8	0.901	0.908	0.901	0.996	0.904	0.901	14.68

Evaluation metrics 0.840 0.998 Set increase [%] 1.85 25.22

Figure 3. The results of the evaluation metrics of the linguistic variable “FWCI 4y” for the Cardinality pattern functions, where the number of

uniquely identified observations exceeded the number of observations from the base approach.

Term	Card. Pattern f	Accuracy	Precision	Recall	Specificity	F1	MCC	Set increase
low	*f3, 0*	0.916	0.922	0.916	0.997	0.919	0.917	26.44
	f3, 0.2	0.905	0.911	0.905	0.997	0.908	0.906	24.14
	f3, 0.4	0.845	0.854	0.845	0.995	0.849	0.850	20.17
	f3, 0.6	0.847	0.860	0.847	0.994	0.853	0.851	12.49
	f4, 0	0.892	0.900	0.892	0.996	0.896	0.893	26.41
	f4, 0.2	0.857	0.864	0.857	0.995	0.860	0.860	24.11
	f4, 0.4	0.842	0.851	0.842	0.995	0.846	0.847	20.15
	f4, 0.6	0.846	0.859	0.846	0.994	0.852	0.850	12.48
	f6, 0.4	0.962	0.963	0.962	0.998	0.962	0.961	1.83
	f6, 0.6	0.924	0.929	0.924	0.997	0.926	0.924	6.74
	f6, 0.8	0.901	0.908	0.901	0.996	0.904	0.901	14.86
high	*f3, 0*	0.877	0.878	0.877	0.989	0.877	0.872	1.11
	f4, 0	0.647	0.645	0.647	0.972	0.645	0.655	1.08

Evaluation metrics 0.645 0.998 Set increase [%] 1.08 26.44

Figure 4. The results of the evaluation metrics of the linguistic variable "FWCI 5y" for the Cardinality pattern functions, where the number of uniquely identified observations exceeded the number of observations from the base approach.

Term	Card. Pattern f	Accuracy	Precision	Recall	Specificity	F1	MCC	Set increase
low	*f3, 0*	0.908	0.914	0.908	0.997	0.911	0.909	30.32
	f3, 0.2	0.873	0.878	0.873	0.996	0.875	0.876	28.11
	f3, 0.4	0.858	0.867	0.858	0.996	0.862	0.862	24.33
	f3, 0.6	0.840	0.854	0.840	0.995	0.847	0.844	17.01
	f4, 0	0.869	0.875	0.869	0.996	0.872	0.872	30.30
	f4, 0.2	0.876	0.881	0.876	0.996	0.878	0.878	28.09
	f4, 0.4	0.855	0.864	0.855	0.995	0.859	0.859	24.31
	f4, 0.6	0.839	0.853	0.839	0.994	0.846	0.843	17.00
	f6, 0.4	0.963	0.964	0.963	0.998	0.963	0.962	1.51
	f6, 0.6	0.926	0.929	0.926	0.997	0.928	0.926	6.30
	f6, 0.8	0.869	0.876	0.869	0.996	0.872	0.872	14.56
mid	*f3, 0*	0.883	0.886	0.883	0.992	0.884	0.880	10.42
	f4, 0	0.651	0.656	0.651	0.980	0.653	0.668	10.42
	f6, 0.4	0.645	0.651	0.645	0.978	0.647	0.661	7.13
	f6, 0.6	0.651	0.656	0.651	0.980	0.653	0.668	10.42
	f6, 0.8	0.651	0.656	0.651	0.980	0.653	0.668	10.42
high	*f3, 0*	0.849	0.850	0.849	0.989	0.849	0.846	13.78
	f4, 0	0.697	0.696	0.697	0.977	0.696	0.702	13.74
	f6, 0.4	0.726	0.727	0.726	0.981	0.726	0.731	7.07
	f6, 0.6	0.701	0.700	0.701	0.979	0.700	0.707	10.65
	f6, 0.8	0.698	0.697	0.698	0.978	0.697	0.704	12.59

Evaluation metrics 0.645 0.998 Set increase [%] 1.51 30.32

Figure 5. The results of the evaluation metrics of the linguistic variable "FWCI NoWindow" for the Cardinality pattern functions, where the number of uniquely identified observations exceeded the number of observations from the base approach.

For the linguistic variable "Team size", the highest set increase was recorded (more than 16% for $f_{3,0,0}$) in the case of term "high" receiving MCC from 0.758 to 0.93. This is the opposite situation to the 4 variables mentioned above. When favoring collaborative publications, in the greatest teams, the classified discipline largely overlaps with the disciplines classified using the base approach. This fact is due that more than 90% of researchers publish only in collaborative teams, rarely having publications written solo (in STEMM disciplines, see full data on collaboration patterns across Europe by discipline in Kwiek 2021 [17]; see also Wagner and Leydesdorff [25]; Wagner [24]; Kamalski and Plume [15]). In the case of term "low" most (except $f_{3,0,0}$) cardinality pattern functions returned Accuracy and MCC scores in the range of 0.585 - 0.606, which means that even on average 40% of authors receive different scientific discipline when favoring publication in the smallest teams (Figure 6.).

Term	Card. Pattern f	Accuracy	Precision	Recall	Specificity	F1	MCC	Set increase
low	*f3, 0*	0.843	0.855	0.843	0.993	0.848	0.845	9.03
	f3, 0.2	0.569	0.590	0.569	0.972	0.579	0.591	0.89
	f4, 0	0.587	0.612	0.587	0.972	0.599	0.606	9.19
	f4, 0.2	0.563	0.587	0.563	0.971	0.574	0.585	1.03
	f6, 0.6	0.564	0.596	0.564	0.971	0.579	0.586	2.87
	f6, 0.8	0.587	0.612	0.587	0.972	0.599	0.605	7.28
mid	*f3, 0*	0.664	0.683	0.664	0.975	0.673	0.672	1.71
	f4, 0	0.607	0.630	0.607	0.973	0.618	0.624	1.99
	f6, 0.6	0.607	0.630	0.607	0.973	0.618	0.624	2.00
	f6, 0.8	0.607	0.630	0.607	0.973	0.618	0.624	1.99
high	*f3, 0*	0.931	0.932	0.931	0.997	0.931	0.930	16.08
	f3, 0.2	0.813	0.816	0.813	0.993	0.814	0.820	9.74
	f4, 0	0.844	0.850	0.844	0.992	0.846	0.845	16.17
	f4, 0.2	0.746	0.753	0.746	0.989	0.749	0.758	9.81
	f6, 0.4	0.923	0.924	0.923	0.996	0.923	0.921	2.46
	f6, 0.6	0.843	0.847	0.843	0.994	0.845	0.847	8.35
	f6, 0.8	0.837	0.841	0.837	0.992	0.838	0.839	13.57

Evaluation metrics 0.563 0.997 Set increase [%] 0.89 16.17

Figure 6. The results of the evaluation metrics of the linguistic variable "Team size" for the Cardinality pattern functions, where the number of uniquely identified observations exceeded the number of observations from the base approach.

The least number of cardinality pattern functions that returned a positive set increase were obtained for the variable "Percentile". The results show that Accuracy and MCC for each function is more than 0.83. In addition, it can be noted that for function f_6, the greater the favoritism of publications in the lower percentile of the journal, the greater the set increase becomes (from 4.36 to 12.3 which is about 3 times additional observations). (Fig. 7). This situation is explained by the fact that in science there are many scientists who tend to submit their papers to journals with a lower prestige (see Blackmore and Kandiko [4]; Franzoni et al.[12]; Starbuck [22]).

Term	Card. Pattern f	Accuracy	Precision	Recall	Specificity	F1	MCC	Set increase
low	*f3, 0*	0.840	0.850	0.840	0.995	0.845	0.845	14.08
	f3, 0.2	0.900	0.908	0.900	0.996	0.903	0.900	1.28
	f4, 0	0.823	0.837	0.823	0.994	0.830	0.829	14.08
	f4, 0.2	0.896	0.904	0.896	0.995	0.900	0.896	1.27
	f6, 0.2	0.837	0.847	0.837	0.996	0.842	0.843	4.36
	f6, 0.4	0.836	0.848	0.836	0.995	0.842	0.841	7.90
	f6, 0.6	0.832	0.844	0.832	0.994	0.838	0.837	10.29
	f6, 0.8	0.830	0.843	0.830	0.994	0.837	0.836	12.30

Evaluation metrics 0.823 0.996 Set increase [%] 1.27 14.08

Figure 7. The results of the evaluation metrics of the linguistic variable "Percentile" for the Cardinality pattern functions, where the number of uniquely identified observations exceeded the number of observations from the base approach.

V. Summary

For the linguistic variables presented above, it can be seen that the application of intelligent counting contributes significantly to increasing the number of unambiguously classified disciplines to authors. For selected Cardinality patter functions it was possible to expect a return that exceeds by 25% the number of results using the base approach (about 30% for Citation and FWCI NoWindow, 26% for FWCI 5y, 25 for FWCI 4y; in all cases for term "low"). One has to wonder, however, whether it is reasonable to use the term "low", if we would like to base our research on the least cited publications and on publications with the lowest FWCI metric. What would be reasonable in this case is to operate on the term "high" and the satisfaction of increasing the number of observations by 13-16% (around 16% for Team Size and 13% for Citation and FWCI NoWindow).

The results presented above provide a basis for further analysis of the presented problem. It is necessary to focus on further unexplored scientific metrics and cardinality pattern functions to examine their influence on the determination of the dominant discipline. Due to the large number (31%) of authors who received assignment to more than one dominant discipline, it would be interesting to consider a multi-label classification variant as an alternative to multi-class classification. Besides discipline, there are other locations where the conceptual apparatus of fuzzy logic can be applied. A dimension that also needs to be explored is the author's dominant country or their full affiliation.

The other approach that may also prove interesting assumes is the use of fuzzy controllers [16]. Appropriate definition of a set of rules for a fuzzy controller can also set a new direction, and thus hypothetically improve the algorithm presented in the classical approach. Besides the scientific database Scopus, there are other databases: Web of Science, OpenAlex (or subsets of these databases) for which the above algorithm can be applied. Therefore, the next step in the work on the algorithm may be to integrate or compare results from many different data sources and applying a voting mechanism to them [14].

Appendix

Table II.
Values of top 10th percentile for membership functions by disciplines

Discipline	Citation	FWCI 4y	FWCI 5y	FWCI NoWindow	Team size	Percentile	Number of publications
AGRI	53	2.36	2.34	2.33	7	53	4,827,009
ARTS	26	2.59	2.53	2.49	3	26	2,144,574
BIOC	74	2.55	2.53	2.46	9	74	8,716,173
BUSI	54	2.98	2.96	2.91	4	54	1,219,375
CENG	66	2.94	2.92	2.88	7	66	2,855,087
CHEM	60	2.67	2.64	2.57	7	60	6,582,601
COMP	45	2.84	2.81	2.68	5	45	2,835,794
DECI	56	2.95	2.95	2.95	4	56	490,649
DENT	45	2.42	2.40	2.38	7	45	405,592
EART	55	2.49	2.47	2.35	7	55	2,800,145
ECON	46	2.75	2.70	2.60	4	46	929,703
ENER	53	3.09	3.08	3.04	7	53	1,611,907
ENGI	40	2.83	2.79	2.62	6	40	9,031,637
ENVI	55	2.79	2.76	2.69	7	55	3,432,530
HEAL	43	2.62	2.59	2.52	7	43	724,899
IMMU	73	2.57	2.54	2.48	10	73	2,037,654
MATE	50	2.80	2.78	2.66	7	50	5,973,127
MATH	35	2.48	2.44	2.32	4	35	3,112,501
MEDI	44	2.19	2.19	2.21	8	44	20,632,131
NEUR	80	2.57	2.55	2.47	9	80	1,690,716
NURS	40	2.59	2.56	2.47	7	40	1,059,202
PHAR	47	2.33	2.31	2.24	8	47	2,337,260
PHYS	51	2.70	2.68	2.50	7	51	7,442,223
PSYC	64	2.64	2.63	2.58	6	64	1,511,213
SOCI	34	2.60	2.56	2.46	4	34	4,492,180
VETE	33	2.46	2.40	2.31	8	33	586,876

Acknowledgment

I gratefully acknowledge the support of my supervisor from Centre for Public Policy Studies Prof. Marek Kwiek. I also gratefully acknowledge the assistance of the International Center for the Studies of Research (ICSR Lab) for research purposes and Kristy James, Senior Data Scientist in the ICSR Lab, for her continuous support.

Author contributions

Lukasz Szymula - Conceptualization, Data curation, Formal Analysis, Methodology, Resources, Software, Validation, Visualization, Writing - original draft, Writing - review & editing.

Krzysztof Dyczkowski - Conceptualization, Formal Analysis, Methodology, Validation, Writing - original draft, Writing - review & editing.

COMPETING INTERESTS

No potential conflict of interest was reported by the author(s).

OPEN ACCESS PRACTICES

Access to the Scopus database was granted through a collaboration between AMU Center for Public Policy Studies and the International Center for the Study of Research (ICSR Lab), Elsevier, established in November 2020 so the dataset is not publicly available. Full results of the study are available on GitHub:
(link: https://github.com/lukaszszy/fedcsis-evaluation-cardinality-pattern-functions-authors-dominant-discipline-classification).

REFERENCES

[1] Abramo, G., Aksnes D. W., D'Angelo C. A., 2020. Comparison of research productivity of Italian and Norwegian professors and universities. Journal of Informetrics, 14(2), 101023.

[2] Albarrán, P., Crespo, J. A., Ortuño, I., Ruiz-Castillo, J., 2011. The skewness of science in 219 sub-fields and a number of aggregates. Scientometrics. Vol. 88(2). 385–397.

[3] Baas, J., Schotten, M., Plume, A., Côté, G., Karimi, R., 2020. Scopus as a curated, high-quality bibliometric data source for academic research in quantitative science studies. Quantitative Science Studies. Vol. 1(1). 377–386. https://doi.org/10.1162/qss_a_00019.

[4] Blackmore, P., & Kandiko, C. B., 2011. Motivation in academic life: A prestige economy. Research in Post-Compulsory Education, 16(4), 399–411.

[5] Boekhout, H., van der Weijden, I., Waltman L., 2022. Gender differences in scientific careers: A large-scale bibliometric analysis . Available from: https://arxiv.org/abs/2106.12624

[6] Carrasco, R., Ruiz-Castillo, J., 2014. The evolution of the scientific productivity of highly productive economists. Economic Inquiry. Vol. 52(1). 1–16.

[7] Cassidy R.S., Larivièrev., 2018. Measuring Research, What Everyone Needs to Know. Oxford University Press.

[8] Chan, H. and Torgler, B., 2020. Gender differences in performance of top cited scientists by field and country. Scientometrics, 125(3), pp.2421-2447, https://doi.org/10.1007/s11192-020-03733-w.

[9] Daradkeh M, Abualigah L, Atalla S, Mansoor W., 2022. Scientometric Analysis and Classification of Research Using Convolutional Neural Networks: A Case Study in Data Science and Analytics. Electronics; 11(13):2066. https://doi.org/10.3390/electronics11132066

[10] de Moya-Anegón, F., Z. Chinchilla-Rodríguez, B. Vargas-Quesada, E. Corera-Álvarez, F. Munoz-Fernández, A. Gonzalez-Molina, and V. Herrero-Solana., 2007. Coverage Analysis of Scopus: A Journal Metric Approach. Scientometrics 73 (1): 53–78.

[11] Dyczkowski K., 2018. Intelligent Medical Decision Support System Based on Imperfect Information. The Case of Ovarian Tumor Diagnosis. Studies in Computational Intelligence, https://doi.org/10.1007/978-3-319-67005-8.

[12] Franzoni, C., Scellato, G., & Stephan, P., 2011. Changing incentives to publish. Science, 333(6043), 702–703.

[13] Jöns, H., Hoyler, M., 2013. Global geographies of higher education: The perspective of world university rankings. Geoforum, 46, pp.45-59, https://doi.org/10.1016/j.geoforum.2012.12.014.

[14] Kacprzyk, J., 1985. Group decision-making with a fuzzy majority via linguistic quantifiers. Part I: a consensory-like pooling. Cybernetics and Systems, 16(2-3), pp.119-129, https://doi.org/10.1080/01969728508927761.

[15] Kamalski, J., and Plume. A., 2013. Comparative Benchmarking of European and US Research Collaboration and Researchers Mobility: A Report Prepared in Collaboration Between Science Europe and Elsevier's SciVal Analytics. Science Europe, Elsevier.

[16] Kickert, W., Mamdani, E., 1978. Analysis of a fuzzy logic controller. Fuzzy Sets and Systems, 1(1), pp.29-44, https://doi.org/10.1016/0165-0114(78)90030-1.

[17] Kwiek, M., 2021. What large-scale publication and citation data tell us about international research collaboration in Europe: Changing national patterns in global contexts. Studies in Higher Education. Vol. 46(12). 2629–2649.

[18] Kwiek, M., Roszka, W., 2022. Academic vs. biological age in research on academic careers: a large-scale study with implications for scientifically developing systems. Scientometrics, https://doi.org/10.1007/s11192-022-04363-0.

[19] Meen C. K., Seojin N., Fei W., Yongjun Z., 2020. Mapping scientific landscapes in UMLS research: A scientometric review. Journal of the American Medical Informatics Association, 27(10), 1612–1624, https://doi.org/10.1093/jamia/ocaa107

[20] Ruiz-Castillo, J., Costas, R., 2014. The skewness of scientific productivity. Journal of Informetrics. Vol. 8(4). 917–934.

[21] Sood, S.K., Kumar, N. & Saini, M., 2021. Scientometric analysis of literature on distributed vehicular networks : VOSViewer visualization techniques. Artif Intell Rev 54, 6309–6341, https://doi.org/10.1007/s10462-021-09980-4

[22] Starbuck, W. H., 2013. Why and where do academic publish? M@n@gement, 5, 707–718.

[23] Visser, M., van Eck, N. and Waltman, L., 2021. Large-scale comparison of bibliographic data sources: Scopus, Web of Science, Dimensions, Crossref, and Microsoft Academic. Quantitative Science Studies, 2(1), pp.20-41, https://doi.org/10.1162/qss_a_00112.

[24] Wagner, C. S., 2018. The Collaborative Era in Science. Governing the Network. Cham: Palgrave Macmillan.

[25] Wagner, C. S., and L. Leydesdorff., 2005. Network Structure, Self-Organization, and the Growth of International Collaboration in Science. Research Policy 34 (10): 1608–18, https://doi.org/10.1016/j.respol.2005.08.002.

[26] Wygralak M., 2015. Intelligent Counting Under Information Imprecision. Applications to Intelligent Systems and Decision Support. , Studies in Fuzziness and Soft Computing, https://doi.org/10.1007/978-3-642-34685-9.

[27] Zadeh L.A., 1975. The concept of a linguistic variable and its application to approximate reasoning—I. Information Sciences, Volume 8, Issue 3, https://doi.org/10.1007/978-1-4684-2106-4_1.

Position Papers of the 18th Conference on Computer Science and Intelligence Systems pp. 119–126
DOI: 10.15439/2023F2898
ISSN 2300-5963 ACSIS, Vol. 36

Towards a Definition of Complex Software System

Jan Žižka
0009-0007-6483-0037
Faculty of Informatics
Masaryk University
Brno, Czech Republic
Botanická 68a, Brno, 60200
Email: jzi@mail.muni.cz

Bruno Rossi
0000-0002-8659-1520
Faculty of Informatics
Masaryk University
Brno, Czech Republic
Botanická 68a, Brno, 60200
Email: brossi@mail.muni.cz

Tomáš Pitner
0000-0002-2933-2290
Faculty of Informatics
Masaryk University
Brno, Czech Republic
Botanická 68a, Brno, 60200
Email: pitner@muni.cz

***Abstract*—Complex Systems were identified and studied in different fields, such as physics, biology, and economics. These systems exhibit exciting properties such as self-organization, robust order, and emergence. In recent years, software systems displaying behaviors associated with Complex Systems are starting to appear, and these behaviors are showing previously unknown potential (e.g., GPT-based applications). Yet, there is no commonly shared definition of a Complex Software System that can serve as a key reference for academia to support research in the area. In this paper, we adopt the theory-to-research strategy to extract properties of Complex Systems from research in other fields, mapping them to software systems to create a formal definition of a Complex Software System. We support the evolution of the properties through future validation, and we provide examples of the application of the definition. Overall, the definition will allow for a more precise, consistent, and rigorous frame of reference for conducting scientific research on software systems.**

***Index Terms*—Software System, Complex System Theory, Complex Software System**

I. Introduction

COMPLEX Systems manifest multifaceted dependencies and interrelationships with other systems and environments, making them difficult, if not impossible, to model in their entirety [1], [2]. Complex Systems show properties that make them peculiar, such as the property of *emergent behavior*, i.e., behavior deriving from the different parts of a system that cannot easily be determined or forecasted when components are observed in isolation.

Complex Systems theory focuses on understanding and explaining the behavior of Complex Systems formed by interacting components [1]. The theory provides a general framework and a set of methodologies to study the emergent properties and dynamics embedded in Complex Systems. However, there is no agreed precise definition of the term as different authors might have other points of view [3]. Complex Systems have been studied in various fields [4]–[6], for example, in social sciences by exploring the complex interactions of individuals in cities.

The successes of software systems in the past years based on, for example, Neural Networks, such as systems developed by DeepMind [7], or OpenAI [8], bring a new range of software systems to wide attention. The appearance of applications such as AlphaFold or ChatGPT and others in a very short time suggests that many more such systems will rise soon.

The exciting behaviors of these new software systems, such as self-organization and emergence, cannot be explained by inspecting the software implementation they are based on. This range of software systems has specific behaviors correlating with the behaviors of Complex Systems as defined by the Complex Systems theory.

While in Computer Science, complexity is studied in different contexts, such as code complexity and the complexity of algorithms [9], this paper focuses on complexity in the context of *software systems*. Also, many software systems are *socio-technical* systems where humans are part of the system rather than only forming its environment. Our study is interested in pure technical systems where humans are not part of the system but may build the system's environment.

Our work aims to provide a clear definition of a Complex Software System (CSS) based on the *theory-to-research* strategy [10], [11], providing a frame of reference about the properties of such systems in relation to what is postulated by Complex Systems' theory [1].

As the field of Complex Systems is evolving [2], we suggest a framework that will also allow the evolution of the definition and terms. The precise definitions will allow more straightforward and unambiguous communication within academia and will be able to connect to existing and future real-world Complex Software Systems. The definition will also provide boundaries for new research fields with a degree of focus cleared of possible ambiguity due to the lack of definitions.

To summarize, we have the following contribution in this article:

- Setting up a framework under which the definition of a Complex Software System is created;
- Defining a Complex Software System based on reference to general Complex Systems theory;
- Listing examples of the use of such a definition;
- Based on the definition and proposed use, list potential future research directions;

The article is structured as follows. In Section II, we provide basic definitions that are commonly adopted in the context

©2023, PTI

Thematic track: Software Engineering for Cyber-Physical Systems

of software systems when discussing Complex Software Systems, such as System of Systems (SoS), Software Ecosystems (SECO), and Complex Adaptive Systems (CAS). The purpose of the need for a precise definition is discussed in Section III. Section IV discusses the method for creating the definition of a Complex Software System. We select several postulates in Section V to form an initial base for defining a Complex Software System. Section VI provides examples of using such a definition. Future research directions are discussed in Section VII, and conclusions are presented in Section VIII.

II. Basic Definitions

In Software Engineering, several commonly used terms and definitions of software systems discuss how software systems and components can be combined and aggregated. This section lists some of the main definitions and examines their relationship to Complex Software Systems.

System of Systems (SoS) is a collection of independent interacting systems [12]. An SoS has several key properties [13]:

- Operational Independence. Any system part of an SoS is self-standing and can operate even if the whole SoS is disaggregated.
- Managerial Independence. Every single system in an SoS is self-governing.
- Geographic Distribution. SoS are often distributed over geographic regions.
- Evolutionary Development. The existence and development of SoS are under constant change.
- Emergent Behaviour. *"Through the collaboration between the systems in an SoS, a synergism is reached in which the system behavior fulfills a purpose that cannot be achieved by, or attributed to, any of the individual systems."* [13]

The systems which are part of an SoS may also be Complex Systems, or the SoS as a whole may form a Complex System – however, the definition of an SoS does not imply that such a system is a Complex System.

Software Ecosystems (SECO) are *"defined as a set of businesses functioning as a unit and interacting with a shared market for software and services, together with relationships among them"* [14]. A SECO may be composed of Complex Software Systems and is a type of SoS. SECOs are typically socio-technical systems [15], which exhibit Complex System behaviors. In a SECO, introducing new elements can potentially have disruptive effects. SECOs features [16], [17] for example contain and provide:

- Inherited characteristics of natural ecosystems like predation, parasitism, mutualism, commensalism, symbiosis, and amensalism.
- Architectural concepts like interface stability, evolution management, security, and reliability.
- Open source development model.
- Platform for negotiating requirements aligning needs with solutions, components, and portfolios.
- Capability for process innovation.
- Controlled central part for the core of the technology.

Complex Adaptive Systems (CAS) *"are systems that have a large number of components, often called agents, that interact and adapt or learn"* [18], [19]. The field of CAS focuses on the adaptive behavior of Complex Systems.

Software CAS refers to software systems that exhibit emergent behavior and self-organization, similar to Complex Adaptive Systems found in nature. These systems can adapt and evolve based on their interactions with the environment through feedback loops. They involve multiple interacting components or agents that collectively exhibit behavior that cannot be easily predicted from the behavior of individual components [18].

A software project may also be considered to be CAS, as suggested by [20].

A subset of Software CAS are Software Self-Adaptive Systems (SSAS) that focus on the ability of a software system to autonomously adapt and modify the behavior or configuration in response to changing conditions or requirements [21], [22]. SSAS have built-in mechanisms that monitor the system's state, analyze environmental changes, and take actions to maintain or improve system properties at runtime [21].

Software CAS [18], [19]:

- Typically operate far from equilibrium.
- Undergo revisions and improvements.
- Do not conform to classic, equilibrium-based mathematical approaches.
- Continuously adapt through recombination of the building blocks.

We summarize the main characteristics of SoS, SECO, and CAS in Table I.

III. Why there is a need for a definition of a Complex Software System?

We need the definition of a Complex Software System for several reasons. Below we discuss benefits, which are the motivators for the research presented in this article.

Clarity and precision: ensure that the meaning of the term Complex Software System is unambiguous.

Consistency: avoid that a software Complex System is defined differently by different researchers or in other contexts, and ensure that the term's meaning remains consistent over time.

Rigor: provide a framework for scientific research. Scientific definitions are necessary for the development of clear, precise, and consistent scientific concepts and for the advancement of scientific research.

Once the definition has been developed, it can be used in various ways. For example for:

Hypothesis testing: support the development of hypotheses about properties of a Complex Software System and its behaviors so that they can be tested through experiments and observations. For example, empirical methods can be used to verify if a software system fulfills the necessary conditions for forming a Complex Software System.

TABLE I
SoS, SECO, CAS Concepts

	System of Systems (SoS)	Software Ecosystems (SECO)	Complex Adaptive Systems (CAS)
Definition	Collection of independent interacting systems [12]	Collection of software components, applications, and services [14]	Collection of components (agents) that interact and evolve [18], [19]
Focus	Collection of interacting systems	Software and services relationships	Software agents interaction
Emergent Behavior	As systems get larger, emergent behavior is more probable [23]	Limited emergence, the introduction of new elements might have disrupting effects	Emergence in terms of adaptive behavior and self-organization
Examples	Smart Grids Systems	Android Ecosystem	Robotic Swarm

Classification: allow classification or categorization of Complex Software Systems based on their properties. For example, software systems use different paradigms, such as Neural Networks or multi-agent architectures. This can be used to create classifications based on system boundaries, technology, or the form of their implementation.

Comparison: a known Complex Software Systems can be compared based on the definition with other systems to find similar or differing properties. This can serve as grounds for expanding the definition or driving the creation of similar software systems.

Theory development: develop new theories or models.

Definitions are essential for academic research, providing a clear and precise framework for developing hypotheses, conducting experiments, and developing theories. Definitions allow researchers to communicate effectively and to build upon each other's work.

IV. Methodology to Build the Definition

To define a Complex Software System, we adopt the *theory-to-research* strategy, in which there is a continuous cycle between theory and empirical validation [10], [11]:

- we extract from the literature (e.g., [1], [2]) common properties of Complex Systems as have been studied in the different fields;
- as there is no full agreement on all properties in the context of the theory of Complex Systems [3], we discuss the most appropriate in the context of software systems, both according to our view of software systems implementation and deployment and with the aid of further related works [4], [6], [18], [24];
- we map each of the properties to a set of identified *necessary*, *sufficient* and *representative conditions* to the context of software systems;
- we provide an initial application of the definition to showcase the main benefits;

Ladyman [2] defines a Complex System based on reviewing attempts in the literature to characterize a Complex System and compiles a set of necessary conditions to represent complexity. In their work, authors provide conditions that are qualitative and which may not be sufficient for complexity, but they set a basis for a way how complexity can be defined. We suggest using the same method for defining a Complex Software System.

We base the method of writing a formal definition of a Complex Software System on a list of properties of three types that are *necessary*, *sufficient*, and *representative* conditions written in natural language.

The set of properties is the definition of a Complex Software System.

The properties will be assembled in the form of graphical frames followed by a commentary. Different types of properties will be color-coded for clarity in the following way, where *"n"* is an ordered number and *"keyword"* is a word abbreviating the property:

Property n - *"keyword"* - Necessary

A property that is a necessary condition but not sufficient to define a Complex Software System. Any Complex Software System must be fulfilling all properties that are necessary conditions. But fulfilling all such conditions doesn't imply that the Software System is Complex Software System.

Property n - *"keyword"* - Sufficient

A property that is the sufficient condition to define a Complex Software System. If a software system fulfils any single property that is sufficient condition then such a system is a Complex Software System.

Property n - *"keyword"* - Representative

A property describing a typical feature of a Complex System is a representative property. Such property is neither necessary nor sufficient but does describe a commonality among Complex Software Systems.

The nomenclature allows referring specific property such as Pn-N *"keyword"* for a necessary condition property, Pn-S *"keyword"* for sufficient condition property and Pn-R *"keyword"* for a representative property.

The numbering of properties is sequential across all the types. The intention and expectation is that the type of the property may change based on future validation and research and keeping the numbering intact will allow for unambiguous referencing.

V. Initial Definition of a Complex Software System

A Complex Software System may be defined by a set of properties which may be viewed as *necessary*, *sufficient*, and *representative* for a system to exhibit Complex System behaviors. Such properties are generic for any Complex System and are also described in existing publications such as [1], [2]. In this section, we will summarize the basic properties of Complex Systems and put those in the context of software systems, creating a base for the definition of a Complex Software System.

Property 1 - *"components"* - Necessary

A Complex Software System is composed of many components.

All definitions of systems complexity [1], [2], [4], [6] require the system to have many components. In the context of Complex Systems, the word *"many"* refers to the term's qualitative rather than quantitative nature. It would, therefore, be incorrect to attempt to quantify it. For example, a system composed of two Complex Systems with manifold interactions and fulfilling other necessary properties is a Complex System, as well as a system composed of millions of components of a similar type, maybe a Complex System. This property comes directly from the definition of the word "system" [25], [26]. Software systems are typically composed of components. This property is a necessary condition but not sufficient for a software system to exhibit complexity. In software, a component may describe different entities based on view or perspective. It can represent a code module, software package, process, or service. From the perspective of a Complex Software System, only a subset of such representations can serve as components in a Complex Software System as they must possess further attributes discussed in the following paragraphs.

Property 2 - *"communication"* Necessary

The components of a Complex Software System have means of intercommunication.

Communication is an essential condition for a Complex Software System. As Ladyman [2] explains: *"Without interaction, a system merely forms a "soup" of particles which necessarily are independent and have no means of forming patterns, of establishing order."*. Communication through messaging shared data, and interfaces is fundamental in software systems. However, this property is not a sufficient condition for a Complex Software System, as many software systems communicate but lack other necessary properties.

Property 3 - *"similarity"* - Representative

The components in a Complex Software System are similar.

Based on Ladyman [2]: *"For interactions to happen and for pattern and coherence to develop, the elements have to be not only many but also similar in nature."*. From the software systems perspective, for example, a system based on front-end, business logic middle-ware, and back-end database components may not form a Complex Software System. This has fascinating implications for software systems, which may be considered complex. However, this condition is not sufficient to determine a Complex Software System. This property may be necessary, but such a statement cannot be demonstrated and proven with the current knowledge and it is a question if a Software System with dis-similar components may still form a Complex Software System or if the system boundaries would exclude such dis-similar components into system's environment rather than being part of the system itself. It also remains to be defined what precisely *similar* means in the context of software components, and similarity needs to be inspected along with heterogeneity and homogeneity.

Property 4 - *"interaction-change"* - Necessary

The strength of components interactions in a Complex Software System is dynamic and changes over time.

"Most interactions are mediated through some sort of exchange process between nodes (components). In that sense, interaction strength is often related to the quantity of objects exchanged." [1]. The interactions among components have to change over time for a Complex Software System to evolve and create a self-organized clustered structure [1]. The resulting network topology contains information about the nodes' and links' dynamics and formation (Chapter 4.5) [1]. This is a familiar property in software systems studies, for example, in the field of dynamic or adaptive networks [27], [28]. This property is necessary for a Complex Software System from which self-organization and clustering emerge.

Property 5 - *"states"* - Necessary

Components of a Complex Software System are characterized by states.

Complex Software Systems are systems that evolve. An algorithmic description of evolution (Chapter 5) [1] is based on the fact that the system has states, and the evolution forms a path through states from time t to time $t+1$. Therefore the existence of states is necessary to create a Complex Software System. The notion of states in software systems is among the basic concepts of any information systems [29]. However, the existence of states is not a sufficient condition for Complex Software Systems.

Property 6 - *"co-evolution"* - Representative

The intercommunication and states of components in a Complex Software System are not independent but co-evolve.

As discussed in (Chapter 1.5) [2] *"Complex systems are characterized by the fact that states and interactions are often not independent but evolve together by mutually influencing each other; states and interactions co-evolve."*. The Complex Systems are characterized by co-evolutionary dynamics (Chapter 4.8) [1]. From a software systems perspective, this can be represented, for example, by adaptive network models [28], which are known to exhibit such co-evolutionary dynamics [30], [31]. The co-evolutionary algorithms may also be used to solve complex software problems [32].

Property 7 - *"context-awareness"* - Representative

A Complex Software System is context-aware.

As shown by Thurner [1], the Complex Systems are often represented by multi-layer networks and *"... for any dynamic process happening on a given network layer, the other layers represent the 'context' ..."*. In other words, such context defines how components on different layers may be influenced. This property is typical for Complex Systems to co-evolve through context dependency and awareness.

Property 8 - *"algorithmicism"* - Representative

A Complex Software System is algorithmic.

Based on Thurner [1], the *"... (Complex Systems) algorithmic nature is a direct consequence of the discrete interactions between interaction networks and states."*. This fits software systems that are naturally algorithmic [29].

The "algorithmicism" may need to be replaced with "intelligence" or "cognition" based on symbolic or sub-symbolic approaches [33], which might be more appropriate from a software systems perspective, when comparing this property to general Complex Systems theory.

Property 9 - *"path-dependency"* - Representative

Complex Software System processes are path-dependent and non-ergodic.

"The Complex Systems are typically governed by path-dependent processes." (Section 2.5) [1]. The process, in a general theory of complex systems, refers to stochastic processes [34]. This further means that processes in complex systems are inherently non-Markovian. It can also be shown that Complex Systems are non-ergodic (for in-depth discussion, see [1]). From the software systems perspective, this means that software system to exhibit such Complex System properties, they must change their boundary conditions as the system evolves.

Property 10 - *"disorder"* - Necessary

A Complex Software System is disordered and out-of-equilibrium.

Ladyman [2] argues that *"...complex systems are precisely those whose order emerges from a disorder rather than being built into them."*. Also, it can be noted that Complex Systems are generally out-of-equilibrium [1], which drives interesting challenges to the concepts of entropy. Although it can be shown [1], [2] that Complex Systems exhibit such properties, it is not obvious how to apply those to software systems.

It must be noted that disorder doesn't imply instability, which might sometimes be associated with the term. In the sense presented by the property, the disorder is related to entropy. For example, the use of GA (Genetic Algorithms) result is seemingly disordered systems when the system is inspected.

Property 11 - *"robust-order"* - Necessary

A Complex Software System exhibits robust order.

The concept of robust order is derived from system disorder. As shown by Ladyman [2] *"... a system consisting of many similar components (elements) which are communicating (interacting) in a disordered way has the potential of forming patterns or structures"*. This refers to self-organization and emergence property. From a software system perspective, this indicates that a Complex Software System shall be composed of similar and at least initially disordered components. This might seem to contradict with P10-N *"disorder"* but *"... although the elements continue to interact in a disordered way, the overall patterns and structures are preserved. A macroscopic level arises out of microscopic interaction, and it is stable"* [2] which Ladyman defines it as a robust order and continues that *"t(T)his kind of robust order is a further necessary condition for a system to be complex"*. Therefore disorder and robust order may co-exist. One example of software systems showing such a property are Artificial Neural Networks (ANN), where initially, input weights of neurons may be initialized with random values and, through learning, such initial disorder forms patterns, structures, or clusters. Also, when examined on a neuron level, ANN will still be disordered.

From a software systems perspective, it will be interesting to study also further run-time uncertainties concerning robust order property.

Property 12 - *"memory"* - Necessary

A Complex Software System has memory.

From Holland [18]: *"A system remembers through the persistence of internal structure"*, Ladyman [2] infer that *"Memory is a straightforward corollary of robust order."*. And Thurner [1] notes that the *"Complex systems often have memory. Information about the past can be stored in nodes (components), if they have a memory, or in the network structure of the various layers."* In such a sense, *memory* refers to the internal self-organized structure of the system. The difference between *memory* and *states* defined by P5-N *"states"* is that *states* represent the system at a specific point in time, but they do not represent history-dependent dynamics stored in the systems *memory*. This property might have various interpretations in software systems, such as a path through imitation-learning [35] or system audit trails. This interesting property might also have yet unknown interpretations in software systems.

Property 13 - *"SoS sufficiency"* - Sufficient

A System of Complex Software Systems is a Complex Software System.

As an intuitive analogy to properties of a Complex Software System – as in Ackoff [12] – it may be possible to show that a system of Complex Software Systems forms a Complex Software System.

This might have exciting implications in practice as once a Complex Software System is created and exists, a new Complex Software System may be formed by creating a system of such systems (SoS).

As the software does not require any material or physical manipulation and software systems can be created relatively quickly, this allows the possible rapid advancement of software-based systems exhibiting Complex System behaviors.

VI. Application of the Definition

A. Unambiguous communication within academia

"Complex Software System" is a widely used term in academia and industry. It refers to a wide range of software systems and viewpoints with a generic notion of a system's complexity. The definition presented in this paper attempts to provide a concrete reference that can be utilized throughout academic discussions to facilitate a common understanding of the term and properties of such a system. Also, the proposed definition framework is intended to extend and refine the definition to support further Complex Software Systems theory development.

The definition of a Complex Software System may be referenced as a whole, or specific properties may be the focus of empirical and other research when studying the properties of software systems. Having a definition of a Complex Software System will bring clarity through academic discussions.

B. Complex Software System categorization

Software systems are open systems [36] with external interactions. The boundary of the system defines what belongs to the system itself and what its surroundings are. The edge of the system may be used for categorization. Many software systems nowadays are socio-technical systems [37] where people are part of the system rather than creating the surroundings and interacting with the system only through the system boundary.

The software systems also interact with humans or are part of machine-to-machine interactions. The software system boundaries can be used as one of the aspects of categorizing types of software systems. Most importantly, this categorization has an essential perspective from Complex Software Systems theory. Most of the socio-technical systems are Complex Systems [37], and the involvement of humans fulfills the necessary conditions presented in Section V.

For example, if we consider the Internet as a Complex Software System, it can be viewed as a socio-technical system. In which case, it fulfills the P4-N "interaction-change" property. The changes are done by human developers, companies, and communities, which interconnect services throughout the Internet. If the boundary of the Internet as a system excludes human actors, the P4-N "interaction-change" might not hold.

The presented properties applied to different boundaries of a software systems can, in this way, provide mechanisms to create a categorization and demonstrate which boundary is allowing the creation of a Complex System and which is not, as they are not fulfilling the necessary conditions defined by the presented properties.

C. Complex Software System modeling

To dive into understanding Complex Software Systems, it will be required to have a model to analyze the properties' effects, how the *necessary* and *sufficient* conditions may be fulfilled or violated, and how *representative* properties may help define a Complex Software System. This can be achieved, for example, by studying existing Complex Systems, as it is done in other fields. However, the challenge is that we might not have access to such systems and, based on the boundary categorization (Section VI-B), some categories of Complex Software Systems might not even exist, for example, pure technical Complex Software Systems, where humans are outside the system boundary. The categorization based on modeling may follow, for example, FTG+PM framework [38], which aims at the categorization of complex cyber-physical systems.

The model can be designed and developed to study Complex Software Systems based on the presented definition and specification of necessary conditions for such a system to exist. Commonality and variability analysis will be required to create such models.

The model will allow experiments to evaluate the assumptions placed by the properties of Complex Software Systems. Understanding underlying principles might show how such software systems may be constructed. The models may also be utilized directly during the process of creation of a complex software system, as, for example, suggested by [39] with SDD (Simulation Driven Development) to tackle inherent system complexity. Modeling ASA (Adaptive Software Architectures)

[40] might be another way to direct the creation of Complex Software Systems.

With models based on the defined Complex System properties, it may be, for example, demonstrated that P13-S "SoS sufficiency" is a *sufficient* condition.

Creating models of different categories of Complex Software Systems is another research path we will follow.

VII. Future research

Creation of the definition of a Complex Software System and a framework for describing the definition will open doors for several research areas:

- Search for and refining Complex Software System properties;
- Exploring categories of Complex Software Systems;
- Creation of Complex Software System models;
- Search for underlying principles of Complex Software Systems;

The framework discussed in Section IV provides means of extending and refining the definition based on future research in the field of Complex Software Systems. The properties may be updated or expanded as new information becomes available. This will require empirical validation of hypotheses and possibly rejecting null hypotheses posed by the definition. The validation is expected to trigger a refinement cycle for the theory, as defined by theory-to-research strategy [10]. The validation may be based on case studies of existing software systems or experimental research based on simulations and modeling.

The list of the initial 13 properties harvested from studies of Complex Systems in other fields may not always have a precise mapping to software systems. Some properties that define necessary conditions will require further research to understand what they can indicate in the context of software systems. Especially P7-R "context-awareness", P9-R "path-dependency", P10-N "disorder", P12-N "memory" or P13-S "SoS sufficiency".

The categorization, discussed in Section VI-B, based on the definition, might provide additional research topics while helping to uncover new underlying general principles of Complex Software Systems. Inspecting academic discussions, aided by systematic literature reviews, surveys, or questionnaires, will facilitate such categorization.

Our research aims at technical systems, excluding socio-technical systems. This constraint, however, might prove to be challenging to isolate, and it is clear that the cyber-physical and socio-technical systems will be encountered during further research where the discrete and continuous aspects of Software Systems co-exist. This will require clearly defining and distinguishing the system boundaries to tackle encountered involvement of non-technical aspects.

System Complexity is viewed as beneficial property of Software Systems exhibiting interesting properties, therefore it is not in the scope of our research to suggest means of easing or reducing the complexity.

Several Software Systems might be considered complex, such as modern operating systems with their constant struggle against cyber attacks. These systems might not fulfill the criteria for complexity based on the proposed definition. In future research, it will be required to categorize such systems and avoid potential confusion on the term as there are different perspectives through which complexity can be viewed.

The proposed definition suggests one sufficient condition P13-S "SoS sufficiency." In future research, we will identify other potential sufficient conditions to extend the definition. However, this task will be challenging.

VIII. Conclusions

Complex Systems were identified and studied in different fields, such as physics, biology, and economics. These systems exhibit properties such as self-organization, robust order, and emergence. In recent years, software systems started to display behaviors associated with Complex Systems, showing previously unknown potential (e.g., GPT-based applications). However, a commonly shared definition of a Complex Software System is not yet available.

For this reason, in this paper, we have presented a definition of a Complex Software System that can serve as a reference for academia to support future research. The definition is a set of 13 initial *necessary*, *representative*, and *sufficient* conditions for a software system to exhibit Complex System behaviors. The properties were selected from Complex Systems research in other fields and mapped to software systems. We suggested allowing for evolution and refinement of the properties, as the definition can be refined by evaluating the studied properties using empirical methods. We have also provided examples of the use of the definition and discussed further research directions in the area of Complex Software Systems.

An unambiguous definition of a Complex Software System is a stepping stone toward understanding its underlying principles.

Acknowledgement

The work was supported from ERDF/ESF "CyberSecurity, CyberCrime and Critical Information Infrastructures Center of Excellence" (No. CZ.02.1.01/0.0/0.0/16_019/0000822).

References

[1] S. Thurner, R. Hanel, and P. Klimek, *Introduction to the theory of complex systems*. Oxford University Press, 2018.
[2] J. Ladyman, J. Lambert, and K. Wiesner, "What is a complex system?" *European Journal for Philosophy of Science*, vol. 3, no. 1, pp. 33–67, Jan 2013. doi: 10.1007/s13194-012-0056-8
[3] H. Ledford, "Language: Disputed definitions," *Nature*, vol. 455, no. 7216, pp. 1023–1028, Oct 2008. doi: 10.1038/4551023a
[4] M. Mitchell, *Complexity: A guided tour*. Oxford university press, 2009.
[5] G. J. Klir and H. A. Simon, *The architecture of complexity*. Boston, MA: Springer US, 1991, pp. 457–476.
[6] M. M. Waldrop, *Complexity: The emerging science at the edge of order and chaos*. Simon and Schuster, 1993.
[7] [Online]. Available: https://www.deepmind.com/
[8] [Online]. Available: https://openai.com/
[9] J. Van Leeuwen, *Handbook of theoretical computer science (vol. A) algorithms and complexity*. Cambridge, MA, USA: Mit Press, 1991. ISBN 0444880712

[10] R. A. Swanson and T. J. Chermack, *Theory building in applied disciplines*. Berrett-Koehler Publishers, 2013.

[11] P. D. Reynolds, *Primer in theory construction: An A&B classics edition*. Routledge, 2015.

[12] R. L. Ackoff, "Towards a system of systems concepts," *Management science*, vol. 17, no. 11, pp. 661–671, 1971. doi: 10.1287/mnsc.17.11.661

[13] C. B. Nielsen, P. G. Larsen, J. Fitzgerald, J. Woodcock, and J. Peleska, "Systems of systems engineering: basic concepts, model-based techniques, and research directions," *ACM Computing Surveys (CSUR)*, vol. 48, no. 2, pp. 1–41, 2015. doi: 10.1145/2794381

[14] D. G. Messerschmitt, C. Szyperski *et al.*, *Software ecosystem: understanding an indispensable technology and industry*. MIT press Cambridge, 2003, vol. 1.

[15] T. Lima, R. P. dos Santos, and C. Werner, "A survey on socio-technical resources for software ecosystems," in *Proceedings of the 7th International Conference on Management of computational and collective intElligence in Digital EcoSystems*, 2015. doi: 10.1145/2857218.2857230 pp. 72–79.

[16] J. Joshua, D. Alao, S. Okolie, and O. Awodele, "Software ecosystem: Features, benefits and challenges," *International Journal of Advanced Computer Science and Applications*, vol. 4, no. 8, 2013. doi: 10.14569/IJACSA.2013.040833

[17] D. Lettner, F. Angerer, H. Prähofer, and P. Grünbacher, "A case study on software ecosystem characteristics in industrial automation software," in *Proceedings of the 2014 International Conference on Software and System Process*, ser. ICSSP 2014. New York, NY, USA: Association for Computing Machinery, 2014. doi: 10.1145/2600821.2600826. ISBN 9781450327541 pp. 40–49.

[18] J. H. Holland, "Complex adaptive systems," *Daedalus*, vol. 121, no. 1, pp. 17–30, 1992.

[19] ——, "Studying complex adaptive systems," *Journal of systems science and complexity*, vol. 19, pp. 1–8, 2006. doi: 10.1007/s11424-006-0001-z

[20] A. B. Myburgh, "Situational software engineering complex adaptive responses of software development teams," *2014 Federated Conference on Computer Science and Information Systems, FedCSIS 2014*, p. 841 – 850, 2014. doi: 10.15439/2014F196

[21] R. De Lemos, H. Giese, H. A. Müller, M. Shaw, J. Andersson, M. Litoiu, B. Schmerl, G. Tamura, N. M. Villegas, T. Vogel *et al.*, "Software engineering for self-adaptive systems: A second research roadmap," in *Software Engineering for Self-Adaptive Systems II: International Seminar, Dagstuhl Castle, Germany, October 24-29, 2010 Revised Selected and Invited Papers*. Springer, 2013. doi: 10.1007/978-3-642-02161-9_1 pp. 1–32.

[22] F. D. Macías-Escrivá, R. Haber, R. del Toro, and V. Hernandez, "Self-adaptive systems: A survey of current approaches, research challenges and applications," *Expert Systems with Applications*, vol. 40, no. 18, pp. 7267–7279, 2013. doi: 10.1016/j.eswa.2013.07.033

[23] J. S. Osmundson, T. V. Huynh, and G. O. Langford, "Emergent behavior in systems of systems," in *INCOSE International Symposium*, vol. 18, no. 1. Wiley Online Library, 2008. doi: 10.1002/j.2334-5837.2008.tb00900.x pp. 1557–1568.

[24] J. M. Ottino, "Complex systems," *American Institute of Chemical Engineers. AIChE Journal*, vol. 49, no. 2, p. 292, 2003. doi: 10.1002/aic.690490202

[25] Merriam-Webster. System. [Online]. Available: https://www.merriam-webster.com/dictionary/system

[26] O. E. Dictionary. system, n. [Online]. Available: https://www.oed.com/view/Entry/196665

[27] F. Kuhn and R. Oshman, "Dynamic networks: models and algorithms," *ACM SIGACT News*, vol. 42, no. 1, pp. 82–96, 2011. doi: 10.1145/1959045.1959064

[28] A. H. Sayed, "Adaptive networks," *Proceedings of the IEEE*, vol. 102, no. 4, pp. 460–497, 2014. doi: 10.1109/JPROC.2014.2306253

[29] I. Sommerville, *Software Engineering*, ser. Always learning. Pearson, 2016. ISBN 9780133943030

[30] D. Antoniades and C. Dovrolis, "Co-evolutionary dynamics in social networks: A case study of twitter," *Computational Social Networks*, vol. 2, no. 1, pp. 1–21, 2015. doi: 10.1109/SITIS.2014.68

[31] M. Farajtabar, M. Gomez-Rodriguez, Y. Wang, S. Li, H. Zha, and L. Song, "Co-evolutionary dynamics of information diffusion and network structure," in *Proceedings of the 24th International Conference on World Wide Web*, 2015. doi: 10.1145/2740908.2744105 pp. 619–620.

[32] P. B. Myszkowski, M. Laszczyk, and D. Kalinowski, "Co-evolutionary algorithm solving multi-skill resource-constrained project scheduling problem," *Proceedings of the 2017 Federated Conference on Computer Science and Information Systems, FedCSIS 2017*, p. 75 – 82, 2017. doi: 10.15439/2017F318

[33] E. Ilkou and M. Koutraki, "Symbolic vs sub-symbolic ai methods: Friends or enemies?" *CEUR Workshop Proceedings*, vol. 2699, 2020.

[34] S. M. Ross, *Stochastic processes*. John Wiley & Sons, 1995.

[35] O. Vinyals, I. Babuschkin, W. M. Czarnecki, M. Mathieu, A. Dudzik, J. Chung, D. H. Choi, R. Powell, T. Ewalds, P. Georgiev *et al.*, "Grandmaster level in starcraft ii using multi-agent reinforcement learning," *Nature*, vol. 575, no. 7782, pp. 350–354, 2019. doi: 10.1038/s41586-019-1724-z

[36] L. v. Bertalanffy, *General system theory: Foundations, development, applications*. G. Braziller, 1968.

[37] V. Tomic, "A bionic view on complex software systems-and the consequences for digital resilience," Master's thesis, Wien, 2021.

[38] S. Mustafiz, J. Denil, L. Lúcio, and H. Vangheluwe, "The ftg+pm framework for multi-paradigm modelling: An automotive case study," *Proceedings of the 6th International Workshop on Multi-Paradigm Modeling, MPM 2012*, p. 13 – 18, 2012. doi: 10.1145/2508443.2508446

[39] T. Baumann, B. Pfitzinger, and T. Jestadt, "Simulation driven development-validation of requirements in the early design stages of complex systems-the example of the german toll system," *Proceedings of the 2017 Federated Conference on Computer Science and Information Systems, FedCSIS 2017*, p. 1127 – 1134, 2017. doi: 10.15439/2017F133

[40] N.-T. Huynh, M.-T. Segarra, and A. Beugnard, "A development process based on variability modeling for building adaptive software architectures," *Proceedings of the 2016 Federated Conference on Computer Science and Information Systems, FedCSIS 2016*, p. 1715 – 1718, 2016. doi: 10.15439/2016F170

Author Index