35th Annual Workshop of the Swedish Artificial Intelligence Society SAIS 2023

June 12-13, 2023
Karlskrona, Sweden

Editors: Håkan Grahn, Anton Borg, and Martin Boldt
Blekinge Institute of Technology, Sweden

Contents

Message from the Chairs

On behalf of the Organizing Committee, it is our pleasure to present the proceeding of the 35th Annual Workshop of the Swedish Artificial Intelligence Society, SAIS 2023.

The SAIS workshop has since its first edition been a forum for building the Swedish AI research community and nurturing networks across academia and industry. Researchers and practitioners in AI and related disciplines, in Sweden and the rest of the world, have been invited to join us in exchanging knowledge, news, and results on AI-related theory and applications.

This edition of the workshop was hosted by Blekinge Institute of Technology, Karlskrona, Sweden, on June 12–13, 2023. 43 contributions were submitted to SAIS 2022. Among 16 submissions to the full paper track, 10 were accepted and included in the proceedings. This means that the acceptance rate of the full papers was 63%. The rest of the submissions were in the form of extended abstracts, including 14 Ph.D. projects, and 2 industrial applications presentations. In total, 26 extended abstracts were accepted for either oral or poster presentation.

We would like to express our gratitude to all authors who submitted their works, and to presenters and participants who guaranteed the success of the workshop by their active presence. We would also like to thank the program committee and the reviewers, session chairs, and administrators. Finally, our special thanks to the highly reputable keynote speakers Dr. Judith Bütepage, Prof. Diego Calvanese, and Prof. Lars Kai Hansen.

Martin Boldt, Anton Borg, and Håkan Grahn
General and Program Chairs of SAIS 2023

Organization and committees

Organizing Committee
Martin Boldt, Blekinge Institute of Technology (General chair)
Anton Borg, Blekinge Institute of Technology (Program co-chair)
Veselka Boeva, Blekinge Institute of Technology
Håkan Grahn, Blekinge Institute of Technology (Program co-chair)
Maria Hedblom, Jönköping University
Lukas Thode, Blekinge Institute of Technology

Steering Chair
Fredrik Heintz, Linköping University

Program Committee

For Full Papers
Anton Borg, Blekinge Institute of Technology (program co-chair)
Håkan Grahn, Blekinge Institute of Technology (program co-chair)
Veselka Boeva, Blekinge Institute of Technology
Paul Davidsson, Malmö University
Göran Falkman, University of Skövde
Maria Hedblom, Jönköping University
Fredrik Heintz, Linköping University
Anders Holst, RISE
Ulf Johansson, Jönköping University
Lars Karlsson, Örebro University
Niklas Lavesson, Blekinge Institute of Technology
Jacek Malec, Lund University
Slawomir Nowaczyk, Halmstad University
André Tiago Abelho Pereira, Royal Institute of Technology
Kalle Åström, Lund University

For Extended Abstracts
Shahrooz Abghari, Blekinge Institute of Technology
Martin Boldt, Blekinge Institute of Technology
Anton Borg, Blekinge Institute of Technology
Paul Davidsson, Malmö University
Håkan Grahn, Blekinge Institute of Technology
Niklas Lavesson, Blekinge Institute of Technology
Lukas Thode, Blekinge Institute of Technology
Florian Westphal, Jönköping University

Paper session 1: AI in Applications

HINTS: Human-Centered Intelligent Realities

Veronica Sundstedt[1], Veselka Boeva[1], Hans-Jürgen Zepernick[1], Prashant Goswami[1], Abbas Cheddad[1], Kurt Tutschku[1], Håkan Grahn[1], Emiliano Casalicchio[1], Markus Fiedler[2], Emilia Mendes[1], Shahrooz Abghari[1], Yan Hu[1], Valeria Garro[1], Thi My Chinh Chu[1], Lars Lundberg[1], and Patrik Arlos[1]

***Abstract*— During the last decade, we have witnessed a rapid development of extended reality (XR) technologies such as augmented reality (AR) and virtual reality (VR). Further, there have been tremendous advancements in artificial intelligence (AI) and machine learning (ML). These two trends will have a significant impact on future digital societies. The vision of an immersive, ubiquitous, and intelligent virtual space opens up new opportunities for creating an enhanced digital world in which the users are at the center of the development process, so-called *intelligent realities* (IRs).**

The "Human-Centered Intelligent Realities" (HINTS) profile project will develop concepts, principles, methods, algorithms, and tools for human-centered IRs, thus leading the way for future immersive, user-aware, and intelligent interactive digital environments. The HINTS project is centered around an ecosystem combining XR and communication paradigms to form novel intelligent digital systems.

HINTS will provide users with new ways to understand, collaborate with, and control digital systems. These novel ways will be based on visual and data-driven platforms which enable tangible, immersive cognitive interactions within real and virtual realities. Thus, exploiting digital systems in a more efficient, effective, engaging, and resource-aware condition. Moreover, the systems will be equipped with cognitive features based on AI and ML, which allow users to engage with digital realities and data in novel forms. This paper describes the HINTS profile project and its initial results.

I. INTRODUCTION

Nowadays, digitalization is an inseparable part of everyone's lives, e.g., how we work, interact, and spend our free time. During the last decade, we have seen that technologies such as augmented reality (AR), virtual reality (VR), and mixed reality (MR), all encompassed in extended reality (XR), have emerged and become available to a broader audience. There have also been numerous discussions on future immersive realities. This development has and will also be propelled by the increase in home work due to the COVID-19 pandemic and sustainability challenges. Consequently, in 2020 the XR market was valued at USD 25.84 billion and is expected to reach USD 397.81 billion by 2026 [1]. Further, there has been tremendous development in artificial intelligence (AI) and machine learning (ML), fueled by vast amounts of data, powerful hardware, and algorithm advances. Both these trends will have a significant impact on future digital societies.

Since 2021, there has been increased focus and discussions on the metaverse, the vision of an immersive and ubiquitous virtual space. It has been proposed to open new opportunities for creating an enhanced human-centered digital world. In such environments, users could experience and interact in a persistent alternative reality that can be modified and customized according to their preferences.

In the interplay of human-computer interaction (HCI), using novel immersive interaction and visualization technologies, a current addition is the dimension of AI. Based on recent developments, it is expected that the future of immersive environments cannot be envisioned without AI. In the last couple of years, new conferences have appeared, such as the IEEE International Conference on Intelligent Reality (ICIR) [2], first organized in 2021, and the Visualization Meets AI [3] workshop in 2020.

AI can enrich the user experience (UX) by making smarter and personalized human-centered choices in future intelligent realities (IRs). However, at the core is still the choice-making human. Therefore, novel interactions and visual analytics techniques in such settings have the potential to influence and aid the entire decision-making process hugely. Furthermore, this process is largely becoming bidirectional, wherein the system learns and predicts based on the user's interests and where the user seeks to guide and play a more active role in personalizing novel immersive environments for interaction.

Digitalization and demands for a more sustainable society will change how we live, develop new products and services, and do business. Enablers have been, e.g., software-intensive systems, mobile communication, and powerful computers. New drivers and enablers have emerged when taking the steps into the next generations of digital societies. Some current trends that will shape future digital societies are:

- users will be able to switch between IRs, expecting seamless and high-quality experiences,
- users expect to interact in visually intuitive ways using XR techniques and new realities,
- data are produced at an increasing pace by vast numbers of heterogeneous sensors and devices,
- there is a tremendous development in AI and ML, and
- information processing is virtualized and seamlessly transmitted across cloud, fog, and edge services.

The remainder of the paper is organized as follows. In Section II, we introduce the background of the HINTS project in light of recent similar projects in Sweden and internationally. Next, Section III discusses our project's objectives and research focus and the novelty it brings to the state-of-the-art. The theme structure of HINTS is presented in Section IV followed by the initial results obtained in

[1]Faculty of Computing, Department of Computer Science, Blekinge Institute of Technology, Sweden veronica.sundstedt at bth.se

[2]Faculty of Computing, Department of Technology and Aesthetics, Blekinge Institute of Technology, Sweden

Section V. Finally, we conclude this paper with a summarised discussion (Section VI) and conclusions (Section VII).

II. BACKGROUND AND RELATED WORK

Computer graphics (CG), visualization, and interactive media have been used for decades in, e.g., computer games and the visualization of complex data. However, over the last decade, we have seen tremendous development in VR, AR, and MR interaction techniques within the umbrella term XR. These technologies initially became available to a broad public audience through commercial alternatives such as *Microsoft Hololens* and *Oculus Rift*. However, there are now more recent headsets available that, for example, include technologies such as eye tracker foveated rendering (ETFR), such as *Meta Quest Pro*.

Further, gestures, gaze, and eye tracking provide additional interaction capabilities. Another significant development in the last decade is the advancement and use of AI and ML, which have impacted and revolutionized many areas, including recommender systems, language and text analysis, translation, computer vision, and autonomous systems. Large amounts of data, powerful hardware, and algorithm development enable this progress. With the explosion of visual content and distributed visual sensor networks, developing novel systems for interaction, visual analysis, scalable communication and computation, and secure and integrity-preserving data handling is required more than ever.

The advancement of CG techniques has substantially improved XR instruments. Those can further be revolutionized by embedding AI to enable XR to communicate and interact effectively with users [4]. This will lead to the development of novel IR systems that can understand and adapt to human behavior and effectively navigate complex user-system interactions. Including intelligence in XR can be helpful for different applications, such as cancer detection [5], gaming [6], advanced visualization [7], driver training [8], and medical training [9]. In addition, such IRs will, in return, support building reliable and transparent AI systems by serving as an extended learning environment for AI ensuring data diversity and representativeness. Training AI using real-world data can be difficult since complicated aspects of human behavior, physical phenomena, and robot dynamics can be challenging to precisely capture in the real world and often require considerable infrastructure costs and manpower [10], [11], [4]. Hence, IR systems can be applied to simulate various cost-intensive, challenging, and potentially dangerous scenarios to enrich real-world data and avoid and mitigate biases.

The *Human-Centered Intelligent Realities* (HINTS) profile project, presented in this paper, is in line with the recent trends in European research advances integrating XR and AI to design intelligent solutions that benefit humans. For example, in the framework of the recently finished Iv4XR H2020 project, a novel verification and validation technology for XR systems based on techniques from AI is developed to provide learning and reasoning over a virtual world. Another H2020 project, EXPERIENCE, has combined VR and AI to explore novel diagnoses and treatment of affective disorders commonly associated with altered multi-sensory perception like depression, anxiety, and eating disorders. The EO4EU is also a European Commission-funded innovation project that aims to make Earth observation data accessible to users through next-generation tools. The Horizon Europe TRANSMIXR project relies on the maturity of XR and AI with the goal of creating a range of human-centric tools for remote content production and consumption via social VR. The SUN is also a social and human-centered project funded by the Horizon Europe programme that aims at investigating and developing XR solutions that integrate the physical and the virtual world in a convincing way from a human and social perspective. The XR2Learn is a Horizon Europe project aiming to combine human-machine interactions with real, mixed, augmented, and virtual environments for the creation of human-centric XR applications in education.

On a Swedish national level, the University of Skövde has a research profile named VF-KDO (Virtual Factory with Knowledge-Driven Optimization) focusing on building and optimizing digital models for the future's production facilities. VF-DKO also uses advanced data mining and interactive visual analytics to extract decision-support knowledge. The Wallenberg AI, Autonomous Systems and Software Program (WASP) is also the largest research initiative in Sweden with a focus relevant to the HINTS profile project.

III. THE HUMAN-CENTERED INTELLIGENT REALITIES PROFILE PROJECT

The HINTS profile builds upon existing experiences and competencies in the areas above and combines these in a synergetic manner to develop novel *human-centered intelligent realities*. The research profile, HINTS, is hosted by the Department of Computer Science (DIDA) at Blekinge Institute of Technology (BTH). HINTS is a six-year project, partly funded by the Knowledge Foundation, which started on the first of September 2022. The profile is led by an established team of researchers at the department with interdisciplinary expertise representing the human-centered IRs perspective. The work is carried out in collaboration with members from initially six external industrial partners: Blackdrop Interactive, Ericsson, IKEA Marketing & Communication, NODA Intelligent Systems, Spotify, and Virotea. The HINTS profile contributes to the strategic direction of BTH and the recruitment of key new personnel in the department.

The project seeks scientific breakthroughs in five interrelated strategic research areas of human-centered IRs: novel experience assessment methodologies, novel environments and interaction techniques, visual analytics, adaptive and distributed AI, and networking. The needs of our industrial partners are grouped into seven industrial challenges. Based on these challenges, five research themes, see Figure 1, are defined with their core research questions, and together, they will address the overall aim of HINTS, i.e., developing *human-centered IRs*.

A. Objective of Research Profile

The *unique* potential of the HINTS research profile is the combined and necessary competence in experience assessment, visual and interactive computing, visualization and visual analytics, ML and data analytics, and cloud-to-edge computing. Furthermore, being the core part of the strategy at BTH, i.e. *digitalization and sustainability*, the vision and commitment of HINTS are to establish an *internationally recognized competence center focusing on human-centered IRs*. A step towards that vision is to establish a competence center at the national level.

The research profile builds upon previous successful projects at BTH. The ViaTecH Synergy project, funded by the Knowledge Foundation (2017-2022), is the main stepping stone for the profile application, where we continue the developments from the last four years to solve challenges for future digital societies.

The proposed profile will also build on the competence developed in the research profile BigData@BTH, funded by the Knowledge Foundation (2014-2020), from where we will utilize the competence in data analytics and ML. Finally, we will build upon the competence developed within the EU projects Bonseyes and BonsApps, which both aim at the more rapid engineering and deeper integration of AI into edge or user equipment, as also foreseen by the HINTS project. Examples of expected results and outcomes include:

- Development of research area excellence, industrial impact, and business value.
- Establishment of a nationally leading and internationally recognized research environment.
- Competence development at BTH within the profile area, i.e., human-centered IRs.
- Development of new and current educational program improvement that reflects the knowledge and competence gained from the profile project.
- Establishment of a stronger and more intimate co-production with existing and new industry partners to ensure successful conditions for future cooperation and development.
- Development of methods and design principles for industrial applications and use cases, validated through demonstrators with industry partners, e.g., tools, support systems, prototypes, proof-of-concept applications, and other artifacts.

B. Overall Objective and Research Question

As outlined in the Background and Related Work in Section II, there are several challenges to address to design and develop novel human-centered IRs for future digital societies. Hence, the overall objective of this research profile is summarized as follows:

- *In co-production with industrial partners and society, develop concepts, principles, methods, algorithms, and tools for human-centered IRs to lead the way for future immersive, user-aware, and smart interactive digital environments.*

In order to pursue the proposed objective, HINTS will investigate the following main research question:

- *How shall we design effective, efficient, and distributed analytical and computational methods for human-centered IRs?*

The main research question is broken down into sub-questions for each research theme. It is important to note that collaboration and interaction between the themes are necessary to address the research questions.

C. Profile Connection and Benefit to Education

The higher education at BTH will benefit strategically from the HINTS profile in several ways. Recently, BTH decided to centralize game-related education at Campus Karlskrona while opening up a new educational media technology track on digital and immersive experiences hosted by the Department of Technology and Aesthetics (DITE) at Campus Karlshamn. Furthermore, HINTS will be a crucial enabler when reshaping the Computer Science-hosted education programs in game technology. Hence, BTH has strategically decided to update and reshape three new game technology programs at BTH, where the theoretical and technical foundations of graphics programming will be a fundamental part.

The computer game industry has long been one of Sweden's fastest-growing industries. However, studies have shown that the supply of skills in the computer games industry needs to be improved and that skills are recruited abroad. It is vital that we can provide education to satisfy the needs of the industry and that both societal and technical research on computer games and interactive techniques should be strengthened [12]. The primary market for students who have completed the program is professional roles in the gaming industry, focusing on game technical implementation and/or design.

However, the increasing interests in visual and data-driven computing and human-centered IRs in several other industries also emphasize the need for broader knowledge that can be applied to digital activities outside the gaming industry. Examples are companies in media production, health, advertising, visualization, architecture, film, training & simulation, and digital vehicles.

As education at DITE focuses on sound, visual, and immersive experiences in both digital and physical environments, the immersive experiences designed by media technology students will provide a promising basis for two-way cross-campus cooperation with HINTS researchers, namely evaluating innovative digital and immersive experiences, and test innovations made within HINTS. Further, HINTS will support the development of advanced-level courses in our newly started (2019) 5-year engineering program "AI and machine learning".

The team connected to HINTS is also responsible for supervising many MSc thesis projects in computer science and most of the BTH 5-year engineering program degree projects, "Game and software engineering". Thus, HINTS will contribute significantly to the strategic development of

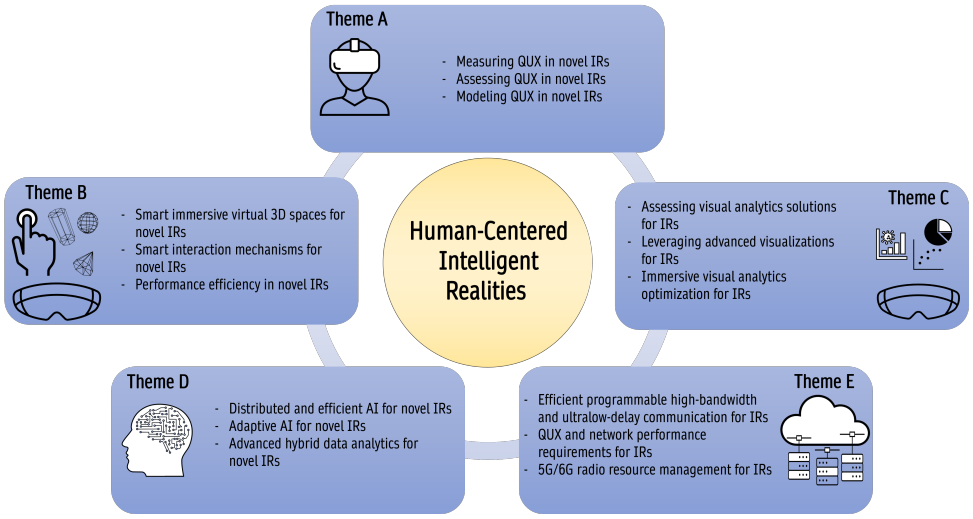

Fig. 1. Overview of the research themes in HINTS.

education at BTH and strengthen the connection between education and research and between departments.

IV. RESEARCH THEMES

In this profile, we will primarily focus on the following aspects organized into five research themes:

A. Theme A: Novel Experience Assessment Methodologies for Intelligent Realities

Novel IR approaches capturing the entire virtual-reality continuum of networked virtual worlds are at the center of Theme A. As the end-user is the final judge of the quality of such applications, it is of utmost importance to understand the behavior and intent of humans in novel IRs as a solid foundation to successfully develop and implement IRs.

Theme A focuses on measuring, assessing, and modeling UX in IRs. These components will lead to an ecosystem of novel experience assessment methodologies ranging from controlled subjective tests to user studies in real environments. This ecosystem goes well beyond the traditional quality of experience (QoE) [13] research of the telecommunications community but also embraces human-centered methods of the UX [14] research from the human-computer interaction community to support the field of quality of user experience (QUX).

QUX, covering QoE and UX, resorts to the concepts of eudaimonia [15]. Eudaimonia aims to understand one's well-being not as an "outcome or end state" but as a "process of fulfilling one's virtuous potentials and living as one was inherently intended to live" [15]. It goes beyond the hedonic and pragmatic usage aspects of interactive applications as it targets the long-term perceived value and usefulness of applications beyond the short-term experience when using an application. Eudaimonia has also been discussed within the context of UX, e.g., [16]. However, to date, only the recent work [17] proposed a multidimensional construct to measure eudaimonia, complementing both QoE and UX. In Theme A, to facilitate the measuring, assessing, and modeling of QUX and its personalization in novel IRs as a formal framework to reveal humans' behavior and intent, we aim at solving the essential challenges associated with immersive computing taking a human-centric approach.

The work toward QUX will reveal an understanding of the user's intent and maneuvering in IRs, which is essential in developing novel human-centered systems that offer the user more interactive and personalized experiences. The development of technical solutions will be grounded on the human-centric approach of the QUX method to include mechanisms of the human visual system, locomotion and navigation, self-motion perception, multimodal and cross-modal interaction and perception, quality perception, cybersickness, presence, task performance, and other human factors. The results from the QUX assessment of human-centric solutions will be used as input to machine learning models to investigate which aspects can be predictors of QUX. Such results will help understand human behavior when using immersive solutions and to identify objective perceptual quality assessment aspects that are relevant within the context of IRs.

B. Theme B: Novel Environments and Interaction Techniques for Intelligent Realities

Theme B aims to investigate the next-generation, novel interaction techniques in XR-based human-centered intelligent 3D virtual spaces or IRs [18]. In these new smart spaces, each user can create personalized virtual environments and interfaces and choose from various modes to interact with them. These spaces are continually enriched by advanced interaction technologies such as haptics, eye-tracking, etc., [19] also allowing multiple users to interact with each other, thereby giving new opportunities for collaboration within the same virtual space (at similar or different locations). On the one hand, Theme B directly interacts with the visualization of these user spaces. On the other, it is itself under evaluation for improvement on the novelty of

experiences. With the added dimension of AI, the system can acquire the capability to predict and personalize the visual content and interaction techniques, taking both the explicit and the tacit hints on user requirements and preferences. Theme B aims to construct the future's intelligent, responsive AI-guided 3D environments and develop the next-generation interaction techniques for single-/multi-users in these virtual environments. A key novelty in these futuristic virtual spaces would be able to adapt to the user's intentions on both these fronts while maintaining efficient performance. This human-machine interaction would thus thrive on the bidirectional feedback between both components.

C. Theme C: Visual Analytics for Intelligent Realities

In the era of the Internet of Things (IoT), large-scale, dynamic, and heterogeneous datasets and the everlasting thrust for exploiting the wealth of information coming out of that, a high-level information abstraction from complex data has become a challenge to existing systems. Therefore, it is paramount to augment human capital using machine intelligence to assist decision-makers in helping them make sense of large quantities of data and use their time effectively. Visual analytics combines our visual intelligence and analysis techniques with visual technology to get relevant information from data [20]. It aims at managing a large amount of data from various sources. It requires a blend of computational analyses and visualizations that can facilitate the understanding and monitoring of complex processes (e.g., machine learning-based 3D immersive and interactive models, unpopulated aerial vehicle vision, and self-driving cars). Current visual analytics systems face multifaceted challenges in adapting to the advancements made by digital realities. In these immersive visual analytics, human-machine (computing intelligence) partnerships will continue to evolve, purposely helping augment cognition in users to answer complex questions. In a nutshell, Theme C revolves around devising efficient (e.g., user non-distractive and comfortable experience) and immersive mechanisms to present and communicate data of complex systems (e.g., analytics of large-scale and multi-source visual data). Theme C's central focus is deep visual analytics which could be transferable to digital realities.

D. Theme D: Adaptive and Distributed AI for Intelligent Realities

Theme D focuses on studying novel resource-efficient adaptive and distributed AI/ML approaches to equip immersive systems with distributed perception and intelligence. Many applications of such immersive systems conferred with intelligence can be identified today, e.g., autonomous cars, sports activities/exercise, gaming applications, advanced visualization methods, smart homes, and many others [4]. These applications require new robust and adaptive AI models that can be run on smart interactive devices with limited power and storage [21], [22]. The models must employ efficient learning and evaluation algorithms capable of dealing with multi-source information varying over time and additionally understanding and predicting user intention and being able to adapt to it. Theme D's main interests are intelligent data-centered solutions focusing on interaction, explainability, and adaptivity, leveraging a wide range of techniques from distributed ML, frequent pattern mining, edge-based AI, continuous learning and domain adaptation, data integration, and analysis. For instance, multi-view data mining algorithms [23] can be used to analyze the heterogeneous data generated by immersive systems and detect user behavioral patterns. These patterns can facilitate the understanding and interpretation of user performance and can be further used to adapt and improve the user behavioral models and experience assessment metrics. Another ambition is to develop novel hybrid immersive analytical techniques that can support reasoning and decision-making in complex data exploration scenarios [24]. AI-assisted visual and advanced data analytics can be combined and used to provide multi-layered analysis of potential problems and direct users to specific points that need attention.

E. Theme E: Networking for Intelligent Realities

IRs will be typically distributed systems. The storage of data and computation will be done at multiple locations to take advantage of remote accelerations hardware, GPUs, or protected data only available at places. IRs will exchange significant amounts of data, often with complex structures requiring highly efficient and resource-aware hardware. Hence, the data transport for IRs is expected to demand, in general, a) ultra-low delays (because of the immersiveness), and even for extensive data, b) no or extremely small delay variation, e.g., aiming at the predictability for starting computation, and c) a network with rapid adaptation towards the efficiency needs for IRs such as enabling the very high-bandwidth data flows to increase the efficiency of the computing resources even on very short notice. The latter task is challenged by the inevitable and potentially long end-to-end delay between two control and processing elements, e.g., between a client on a device and a server or acceleration hardware in the cloud. The delay might block network-application interactions that are across multiple layers. Fast network adaption or data flow management can be executed efficiently within the network layer by "in-network" mechanisms. SDN technologies, such as OpenFlow, P4, or eBPF, have recently demonstrated such capabilities and are now components of modern communication networks 5G and future 6G systems [25].

Each of the three general network performance objectives ((a) - (c)) for IRs is difficult to tackle. Hence, any combination of them is even more demanding. However, networking mechanisms that address them jointly will have the necessary capabilities for IRs. The pragmatic formulation of the objective is to provide the "right data at the right time and at the right place".

The pragmatic notion can be translated into three subparts of a networking system model in HINTS, as shown in Figure 2. The model is initial and continuously refined. The model's first subpart (Part 1) is the detailed specification of QUX and network performance requirements. This part

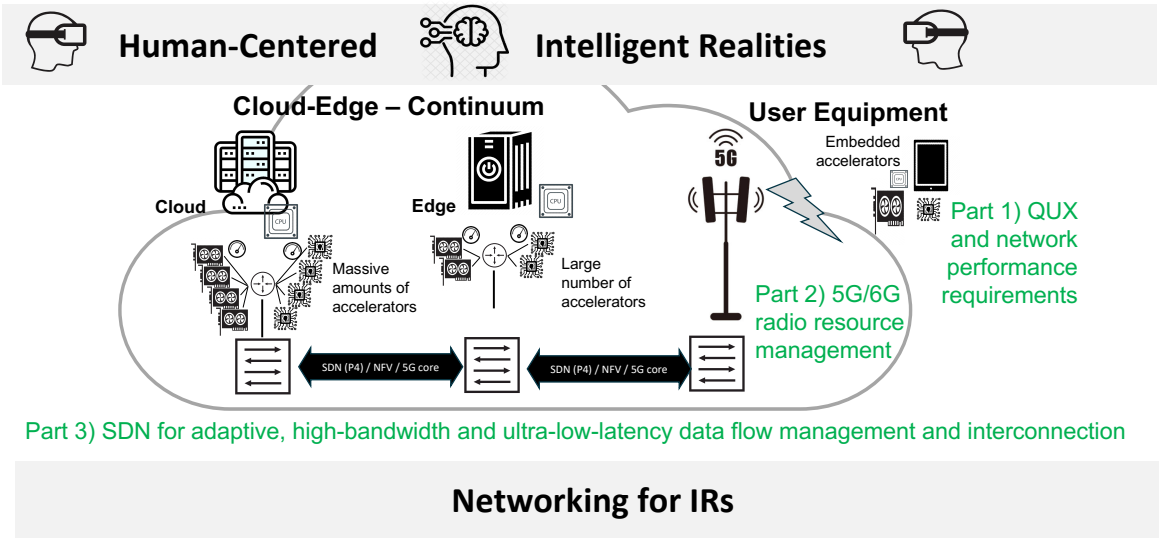

Fig. 2. System model for networking in HINTS.

aims at a calculation model to relate both performance views, i.e., the one from the user experience and the one from the network performance. A significant challenge is the use of new assessment concepts (cf. Theme A) and the relations of perceived immersiveness and QUX with network performance. The second subpart of the model (Part 2) is the minimization of the latency in accessing the network for 5G and 6G by applying sophisticated radio resource management. The third subpart of the model (Part 3) is the use of SDN technologies for adaptive, high-bandwidth, and ultra-low-latency data flows management and interconnection services in cloud and edge computing environments for IRs. For example, Part 3 on SDN-enabled management may use the relationship model from Part 1 to control its in-network mechanisms, i.e., application AI, to manage the network.

In the future, the HINTS networking model needs to address more objectives for enabling sustained efficiency for IRs. These objectives comprise factors such as the reduction of energy consumption in data transmission as well as in computation, e.g., by enabling data forwarding to more energy-efficient acceleration hardware, which is eventually located in the Cloud-Edge computing continuum. Another future objective is to increase the security of IRs by the expanding use of distributed algorithms, such as Blockchains for secure data exchange in federated learning (FL) [26].

V. INITIAL RESULTS

Apart from the initial works being broken down in more detail, the major developments so far are highlighted below. A new laboratory has been set up to promote collaboration in a complete environment merging strong education programs, research excellence, and industrial and societal collaboration. The laboratory has already started to be used in courses relevant to the area at BTH, which are further exemplified below. A HINTS web page has also been established as part of the project startup phase.

A. Human-Centered Intelligent Realities Laboratory

The Human-Centered Intelligent Realities Laboratory (HCIRL) is a centrally funded strategic infrastructure initiative at BTH aimed at creating a strong research and education environment. This is part of a more considerable BTH strategic investment in infrastructures, where university-wide laboratory facilities are essential. The goal of the laboratory is to have a shared, flexible, creative space for research, education, and activities with industry and society. Expected synergies via the laboratory are foreseen with new project developments. Example activities planned in the laboratory are as follows:

- joint educational and research environment,
- internal and external collaborations,
- basic and advanced level undergraduate and postgraduate courses,
- joint training workshops with industrial partners,
- regional test-bed initiative,
- targeted course moments in education,
- demo room for targeting public and external visits,
- dedicated staff for long-term operation, and
- strategic space for new joint initiatives.

In the laboratory under development, we have the following research equipment and software: i) MR smartglasses (two *HoloLens 2* units), ii) VR headsets (one *Meta Quest 2* unit), iii) AR tablets (one *Samsung Galaxy Tab S7* series), iv) Eye trackers (three *Tobii Eye Tracker 5* units), and v) cleaning devices for AR/VR HMDs (two *CleanBox CX1 (Cone A/Cone B)* units). The lab also has the iMotions biometric research software platform with the incorporated VR Eye Tracking Module. A screen has also been mounted outside the laboratory to disseminate results, invitations, and contact information for research studies and ongoing

Fig. 3. The Human-Centered Intelligent Realities Laboratory: (**left**) outside view and (**right**) CleanBox CX1 units for cleaning the XR headsets.

activities. The HINTS profile[1] brings 12 servers for AI computing and storage. More specifically, HINTS supplies 1) two GPU compute-oriented servers, each with an Nvidia Tesla-T4 GPU, a pair of Intel Xeon 6326 CPUs with 32 cores, 512 GB of RAM, and 12 TB of storage, 2) four general compute servers, each with a pair of Intel Xenon 4314 CPUs with 32 cores, 512 GB of RAM, and 12 TB of storage, 3) six general-purpose nodes, each with 32 TB of storage. In addition to these, HCIRL has a setup of four high-end desktop PCs with various spreads in the combination of CPUs and GPUs. Two systems are based on Intel Core i9-12900K CPU, and the others have AMD Ryzen 9 5950X CPU. The GPUs are Nvidia RTX-3080Ti and AMD Radeon 6900XT. Each desktop is connected to two high-end screens with 1440p@170Hz or 4K@144Hz.

B. Research-Augmented Education

There have been several courses at BTH using the equipment so far, in addition to the research activities. A researcher from the HINTS profile is involved in the PA2570: Behavioural Software Engineering course, which explores human factors in software engineering. As part of the course involvement, the eye tracking technology in the environment is used together with analysis software during a lab assignment. Eye tracking has previously been used in education at BTH, for example, in a Visualization course curriculum [27]. The HoloLens 2 headset is also used by a student in the TE2502: Degree Project in the Master of Science in Engineering course for the final thesis project to evaluate gesture-based interaction in XR.

Networking research topics for IRs (cf. Section IV-E) have also been already included in BTH's education. For example, a degree project in the Master of Science in Telecommunication Systems (course ET2606) addresses the applicability of eBPF for in-network management for IRs. This educational work was reinforced by publishing a book on the programmability and virtualization of 5G and Beyond 5G networks [25]. Researchers from the HINTS profile have also separately been responsible for developing a new course, DV2583: Digital Ethics, as part of the 5-year engineering program "AI and machine learning". The course is crucial, covering issues related to AI and human behavior in the digital society.

[1]The PROMIS project has an identical setup, and resources can be shared among the projects, depending on needs.

C. Initial HINTS Publications

The Theme D researchers involved in HINTS are active in two main areas of the project: data mining and analysis and distributed and adaptive ML. In the first area, the researchers have recently published a comprehensive survey of state-of-the-art intelligent fault detection and diagnosis (FDD) in district heating (DH) systems [28]. The survey analyzes the developments in intelligent FDD for the DH domain, identifies current research gaps and techniques limitations, and supplies recommendations for future studies. These all will put the baseline and boost the planned collaboration research with some industrial partners involved in HINTS, e.g., NODA Intelligent Systems. In the second area, one of the pursued research directions involved studying approaches that bring efficiency and robustness to FL settings, as discussed in [29]. In a paper [30], researchers from Theme D have proposed a novel FL model that tries coping with statistically heterogeneous environments by introducing a group-personalized FL method. Such solutions are in the research focus of some of our industrial partners, e.g., Ericsson. Furthermore, the study published in [26] investigates Blockchains for the secure data exchange for distributed and federated AI/DL learning. Theme B has ongoing conference/journal submissions related to VR, CG, and AI. The research in these fields forms a crucial basis for multiple themes and subprojects in HINTS and reflects on this project's intertwined multidisciplinary aspect.

VI. DISCUSSION

The recent digitization of industrial and social processes generated not only a tsunami of data and information but also made users thinking how they can prevail in their interaction with digital systems and processes. In addition, the current COVID-19 pandemic forced people to find new ways of working, living, and interacting. As a result, they needed

to use digital tools quickly, as recently researched by McKinsey [31]. However, the pandemic has also demonstrated how indispensable digital technologies are, as outlined on the international governmental level [32]. As a result, it became evident that current means to digitize processes and handle digital information are insufficient. Hence, new ways are required to understand, collaborate with, and control future generations of digital systems. HINTS will address such issues using a novel approach founded in visual and data-driven platforms which enable tangible, immersive cognitive interactions within real and virtual realities.

HINTS will address the identified challenges in the related literature as follows: i) concerning immersive computing, an iterative approach will be employed, where three areas will co-create results by exchanging each other's methods and results. These areas are to respectively measure, assess and model the QUX in novel IRs; ii) concerning novel environments and interaction techniques for IRs, HINTS will contribute via three pillars to various aspects of novel IRs, i.e., smart immersive virtual 3D spaces, smart interaction mechanisms, and performance efficiency; iii) about enhancing visual analytics for IRs, it will compare and assess visual analytics solutions, leverage advanced visualizations, and put forward and validate immersive visual analytics optimization; iv) within the context of adaptive and distributed AI for IRs, it will implement three types of solutions: distributed and efficient AI, adaptive AI, and advanced hybrid data analytics; and v) regarding networking for IRs, it aims at tackling the relationship of QUX and network performance, defining a relationship model that can be inverted, i.e., one that use network performance as input and activates network actions in the network if QUX objectives are not met, and investigating SDN-and in-network based flow management mechanisms for high-bandwidth and ultra-low-latency IR services. In addition, novel engineering concepts of 5G and 6G radio resource management for IRs will be researched.

VII. CONCLUSIONS

This paper has introduced the profile project HINTS – "Human-Centered Intelligent Realities". HINTS is centered around an ecosystem combining XR and communication paradigms to form novel intelligent digital systems and builds upon four previously funded projects, two financed by the Knowledge Foundation (the ViaTecH Synergy project and the research profile BigData@BTH) and two EU projects - Bonseyes and BonsApps. HINTS addresses the needs of six industrial partners: Blackdrop Interactive, Ericsson, IKEA Marketing & Communication, NODA Intelligent Systems, Spotify, and Virotea. Such needs were arranged into five interrelated strategic research themes of human-centered IRs, and are as follows: 1) novel experience assessment methodologies, 2) novel environments and interaction techniques, 3) visual analytics, 4) adaptive and distributed AI, and 5) networking.

HINTS aims to be the central Swedish node with a high international impact in human-centered IRs for next-generation digital societies. Finally, the HINTS profile contributes to the strategic direction of BTH (both in research and education) and the recruitment of competent new personnel at the Computer Science department. Being at the center of the BTH strategy towards digitalization, the HINTS focus and its complete environments are built upon strong academic programs, research excellence, and co-production with external partners.

ACKNOWLEDGMENT

This research was funded partly by the Knowledge Foundation, Sweden, through the Human-Centered Intelligent Realities (HINTS) Profile Project (contract 20220068). For more information, please check the HINTS web page.

REFERENCES

[1] Extended Reality (XR) Market - Growth, Trends, COVID-19 Impact, And Forecasts (2023 - 2028). (accessed: 28.03.2023). [Online]. Available: https://www.mordorintelligence.com/industry-reports/extended-reality-xr-market

[2] "IEEE 2^{nd} International Conference on Intelligent Reality," (accessed: 28.03.2023). [Online]. Available: https://icir.ieee.org/

[3] "Visualization Meets AI," (accessed: 28.03.2023). [Online]. Available: https://vismeetsai.github.io/

[4] D. Reiners, M. R. Davahli, W. Karwowski, and C. Cruz-Neira, "The combination of artificial intelligence and extended reality: A systematic review," *Frontiers in Virtual Reality*, vol. 2, p. 721933, 2021.

[5] P.-H. C. Chen, K. Gadepalli, R. MacDonald, Y. Liu, S. Kadowaki, K. Nagpal, T. Kohlberger, J. Dean, G. S. Corrado, J. D. Hipp *et al.*, "An augmented reality microscope with real-time artificial intelligence integration for cancer diagnosis," *Nature medicine*, vol. 25, no. 9, pp. 1453–1457, 2019.

[6] E. Turan and G. Çetin, "Using artificial intelligence for modeling of the realistic animal behaviors in a virtual island," *Computer Standards & Interfaces*, vol. 66, p. 103361, 2019.

[7] A. H. Sadeghi, A. P. Maat, Y. J. Taverne, R. Cornelissen, A.-M. C. Dingemans, A. J. Bogers, and E. A. Mahtab, "Virtual reality and artificial intelligence for 3-dimensional planning of lung segmentectomies," *JTCVS techniques*, vol. 7, pp. 309–321, 2021.

[8] S. Ropelato, F. Zünd, S. Magnenat, M. Menozzi, and R. Sumner, "Adaptive tutoring on a virtual reality driving simulator," *International SERIES on information systems and management in creative emedia (CreMedia)*, vol. 2017, no. 2, pp. 12–17, 2018.

[9] V. Bissonnette, N. Mirchi, N. Ledwos, G. Alsidieri, A. Winkler-Schwartz, and R. F. Del Maestro, "Artificial intelligence distinguishes surgical training levels in a virtual reality spinal task," *The Journal of bone and joint surgery. American volume*, vol. 101, no. 23, p. e127, 2019.

[10] W. Guerra, E. Tal, V. Murali, G. Ryou, and S. Karaman, "Flightgoggles: Photorealistic sensor simulation for perception-driven robotics using photogrammetry and virtual reality," in *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2019, pp. 6941–6948.

[11] A. Amini, I. Gilitschenski, J. Phillips, J. Moseyko, R. Banerjee, S. Karaman, and D. Rus, "Learning robust control policies for end-to-end autonomous driving from data-driven simulation," *IEEE Robotics and Automation Letters*, vol. 5, no. 2, pp. 1143–1150, 2020.

[12] B. Flintberg and J. Nylander, *Kraftsamling Dataspelsbranschen : En rapport om svensk spelindustri*. RISE Rapport, 2023.

[13] K. Brunnström, S. A. Beker, K. De Moor, A. Dooms, S. Egger, M.-N. Garcia, T. Hossfeld, S. Jumisko-Pyykkö, C. Keimel, M.-C. Larabi *et al.*, "Qualinet white paper on definitions of quality of experience," 2013.

[14] T. Zaki and M. N. Islam, "Neurological and physiological measures to evaluate the usability and user-experience (ux) of information systems: A systematic literature review," *Computer Science Review*, vol. 40, p. 100375, 2021.

[15] E. L. Deci and R. M. Ryan, "Hedonia, eudaimonia, and well-being: An introduction," *Journal of happiness studies*, vol. 9, pp. 1–11, 2008.

[16] E. D. Mekler and K. Hornbæk, "Momentary pleasure or lasting meaning? distinguishing eudaimonic and hedonic user experiences," in *Proceedings of the 2016 chi conference on human factors in computing systems*, 2016, pp. 4509–4520.

[17] F. Hammer, S. Egger-Lampl, and S. Möller, "Quality-of-user-experience: a position paper," *Quality and User Experience*, vol. 3, pp. 1–15, 2018.

[18] K. Rook, B. Witt, R. Bailey, J. Geigel, P. Hu, and A. Kothari, "A study of user intent in immersive smart spaces," in *2019 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops)*. IEEE, 2019, pp. 227–232.

[19] D. Navarro, V. Sundstedt, and V. Garro, "Biofeedback methods in entertainment video games: A review of physiological interaction techniques," *Proceedings of the ACM on Human-Computer Interaction*, vol. 5, no. CHI PLAY, pp. 1–32, 2021.

[20] J. Thomas and P. C. Wong, "Visual analytics," *IEEE Computer Graphics and Applications*, vol. 24, no. 5, pp. 20–21, 2004.

[21] F. Zhuang, Z. Qi, K. Duan, D. Xi, Y. Zhu, H. Zhu, H. Xiong, and Q. He, "A comprehensive survey on transfer learning," *Proceedings of the IEEE*, vol. 109, no. 1, pp. 43–76, 2020.

[22] J. Verbraeken, M. Wolting, J. Katzy, J. Kloppenburg, T. Verbelen, and J. S. Rellermeyer, "A survey on distributed machine learning," *Acm computing surveys (csur)*, vol. 53, no. 2, pp. 1–33, 2020.

[23] J. Zhao, X. Xie, X. Xu, and S. Sun, "Multi-view learning overview: Recent progress and new challenges," *Information Fusion*, vol. 38, pp. 43–54, 2017.

[24] M. Cavallo, M. Dolakia, M. Havlena, K. Ocheltree, and M. Podlaseck, "Immersive insights: A hybrid analytics system forcollaborative exploratory data analysis," in *Proceedings of the 25th ACM Symposium on Virtual Reality Software and Technology*, 2019, pp. 1–12.

[25] L. J. Horner, K. Tutschku, A. Fumagalli, and S. Ramanathan, *Virtualizing 5G and Beyond 5G Mobile Networks*. Artech House, 2023.

[26] V. Daliparthi, N. Momen, K. Tutschku, and M. De Prado, "ViSDM: A liquid democracy based visual data marketplace for sovereign crowdsourcing data collection," in *Proceedings of the 2023 European Interdisciplinary Cybersecurity Conference*, 2023.

[27] V. Sundstedt, "A visualisation course in a game development curriculum." in *Eurographics (Education Papers)*, 2016, pp. 9–16.

[28] J. van Dreven, V. Boeva, S. Abghari, H. Grahn, J. Al Koussa, and E. Motoasca, "Intelligent approaches to fault detection and diagnosis in district heating: Current trends, challenges, and opportunities," *Electronics*, vol. 12, no. 6, p. 1448, 2023.

[29] A. A. Al-Saedi, V. Boeva, and E. Casalicchio, "Fedco: Communication-efficient federated learning via clustering optimization," *Future Internet*, vol. 14, no. 12, p. 377, 2022.

[30] A. A. Al-Saedi and V. Boeva, "Group-personalized federated learning for human activity recognition through cluster eccentricity analysis," in *24th International Conference on Engineering Applications and Advanced AI (EANN/EAAAI 2023)*, 2023, accepted.

[31] A. Baig, B. Hall, P. Jenkins, E. Lamarre, and B. McCarthy, "The COVID-19 recovery will be digital: A plan for the first 90 days," (accessed: 28.03.2023). [Online]. Available: https://www.mckinsey.com/business-functions/mckinsey-digital/our-insights/the-covid-19-recovery-will-be-digital-a-plan-for-the-first-90-days

[32] "ITU. Ministerial Roundtable 1: The role of digital technologies during and after the COVID 19 pandemic," (accessed: 28.03.2023). [Online]. Available: https://digital-world.itu.int/ministerial-roundtable-1-the-role-of-digital-technologies-during-and-after-the-covid-19-pandemic/

Fully Convolutional Networks for Dense Water Flow Intensity Prediction in Swedish Catchment Areas

Aleksis Pirinen[1], Olof Mogren[1] and Mårten Västerdal[2]
[1]RISE Research Institutes of Sweden
[2]Department of City Planning and Sustainability, Kungsbacka municipality
{aleksis.pirinen@ri.se, olof.mogren@ri.se, marten.vasterdal@kungsbacka.se}

***Abstract*— Intensifying climate change will lead to more extreme weather events, including heavy rainfall and drought. Accurate streamflow prediction models which are adaptable and robust to new circumstances in a changing climate will be an important source of information for decisions on climate adaptation efforts, especially regarding mitigation of the risks of and damages associated with flooding. In this work we propose a machine learning-based approach for predicting water flow intensities in inland watercourses based on the physical characteristics of the catchment areas, obtained from geospatial data (including elevation and soil maps, as well as satellite imagery), in addition to temporal information about past rainfall quantities and temperature variations. We target the one-day-ahead regime, where a fully convolutional neural network model receives spatio-temporal inputs and predicts the water flow intensity in every coordinate of the spatial input for the subsequent day. To the best of our knowledge, we are the first to tackle the task of *dense* water flow intensity prediction; earlier works have considered the prediction of flow intensities at a sparse set of locations at a time. An extensive set of model evaluations and ablations are performed, which empirically justify our various design choices. Code and preprocessed data have been made publicly available at `https://github.com/aleksispi/fcn-water-flow`.**

I. INTRODUCTION

As climate change intensifies, hydrological conditions will change. This will manifest itself both in the form of water shortages and as flooding in cases of intense precipitation. According to the Swedish Environmental Protection Agency, the climate in Sweden is becoming warmer and wetter [1], and municipalities are encouraged to increase their climate adaptation efforts, especially regarding mitigating the risks of, and damages associated with, flooding [12], [15]. The effects of climate change on rainfall-runoff will be more severe further north [14]. At the same time, the hydrological conditions in Sweden have been severely disturbed during the last two hundred years, with wetlands being drained and natural streams being straightened, which will further increase the effects of extreme weather events.

Hydrological modeling can shed light on the dynamics of water flow and how it is affected by various aspects of the environment. This can in turn allow for making informed decisions about the efficacy of nature-based climate change adaptation techniques such as wetland restoration, urban greening, and soil protection. Traditional hydrological models are based on expert knowledge and physical properties such as the preservation of volume, which have to be specified a priori. These work well for a certain domain if they are properly calibrated, but have difficulties generalizing to wider environmental categories [17]. Statistical data driven modeling, including machine learning (ML), is an alternative which has the potential to become more robust as long as the model can be trained on a large enough dataset with a suitable learning signal. This way, not only can the flow intensity be estimated for any given water course following a heavy precipitation event, but also the response time of the given area, i.e. an estimation of the time lapse from the precipitation event to peak flow. Such information is vital to better understand flood risks and the effects of flood and drought mitigation, as well as general hydrological implications from changes in land use.

In this work we propose an ML-based approach for water flow intensity prediction that leverages the physical characteristics of a catchment area. We target the one-day-ahead regime, where a fully convolutional neural network [11] receives spatio-temporal inputs and predicts the water flow intensity at every coordinate for the subsequent day (the same modeling should however be able to handle other time horizons with minor modifications). Two important novelties of our proposed approach are:

- In addition to temporal data (past rainfall and temperatures), we include spatial data as inputs to the modeling, provided as satellite imagery and several derived GIS layers. This allows the model to build internal representations about relationships between temporal and spatial aspects of the local environment (including land cover, soil depth and moisture, and elevation).
- Using a fully convolutional model, we tackle the task of *dense* water flow intensity prediction, as opposed to only predicting flow intensities for a sparse set of spatial locations. To the best of our knowledge, we are the first to consider the dense prediction task.

The remainder of this paper is organized as follows. In Section II we provide a brief overview of the related work. Then, in Section III, we describe in detail the data we have used for modeling, training and evaluation. In Section IV we explain our approach for tackling the water flow intensity prediction task, and our proposed approach is empirically evaluated against alternative methods in Section V-B. Finally, the paper is concluded in Section VI.

II. RELATED WORK

Water flow prediction (also known as *stream forecasting* or *rainfall-runoff modeling*) for rivers in the U.S. have been modeled using Long Short-Term Memory (LSTM [7]) networks [8], [5], [4], [6]. The modeling follows a traditional setup inspired by earlier physics-based hydrological models such as the U.S. National Water Model (NWM), based on WRF-Hydro [2]. Jia et al. [8] modeled river segments using an LSTM network with graph convolutions. One segment in the river network corresponds to a distance that the water flows during approximately one day. Input features include daily average precipitation, daily average air temperature, date of the year, solar radiation, shade fraction, potential evapotranspiration, elevation, length, slope, and width of each segment. Models were trained using a physics-informed setup where a traditional flow model acted as a teacher for the machine learning model. LSTM networks have also been used for post-processing the output from the NWM [5]. Similar to our work, most of these prior works have focused on next-day predictions. However, there are examples of hourly predictions [6].

Others have also employed convolutional neural networks (CNNs) for stream forecasting [3], [18], [13]. However, in contrast to us, these works do not incorporate spatial data from satellites or GIS, but instead model only the much lower-dimensional data (in single coordinates or very small neighborhoods, not entire areas as in our setup) provided as a feature vector for each time step, similar to the models using LSTMs. More broadly, deep learning has been used for many related tasks, such as groundwater level estimation [21], water quality estimation [19], and rainfall-runoff [10].

While some of the above mentioned works on streamflow estimation include information about the near environment (such as elevation and slope), none of them use detailed spatial information inputs as is proposed in our work. The use of fully convolutional neural networks to encode this information, in combination with traditional inputs such as rainfall and temperature, has the potential of representing more complex relationships and can result in a more detailed view of the near environment. It also enables us to perform *dense* water flow intensity prediction, different to prior works who consider the prediction of flow intensities in a sparse, discrete set of points.

III. DATASET DESCRIPTION

We use data from 12 locations in Sweden, based on where the Swedish Meteorological and Hydrological Institute (SMHI) has stations for measuring weather and water flow data. These locations are Jönköping (Tabergsån), Knislinge (Almaån), Krycklan, Skivarp (Skivarpsån), Skövde (Ösan), Torup (Kilan), Tumba (Saxbroån), Dalbergsån, Degeå, Hässjaån, Lillån, and Lillån-Blekinge (see Figure 1).

In each location we have access to the following spatial data layers (see also Figure 2):

- satellite RGB image (Sentinel-2) from the Land Survey of Sweden (Lantmäteriet), 10m × 10m resolution;

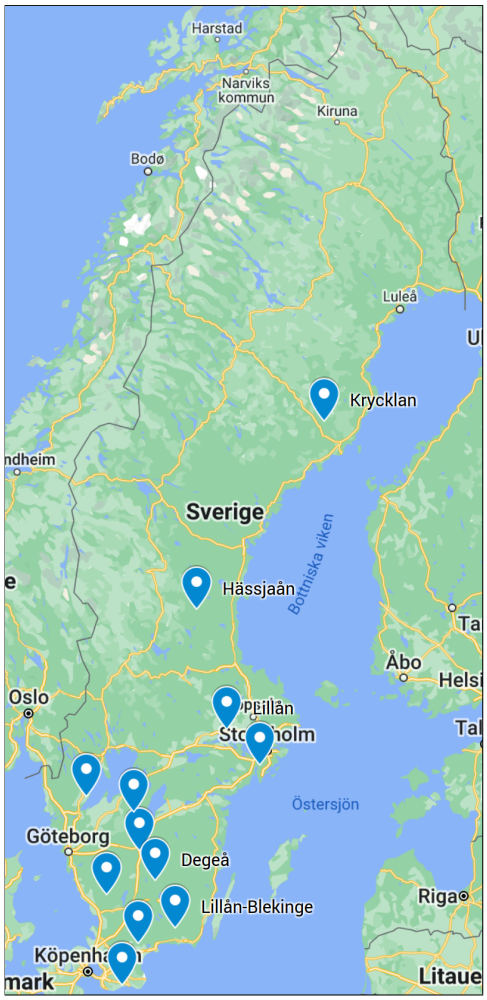

Fig. 1. The data used in this paper comes from 12 locations in Sweden: Jököping (Tabergsån), Knislinge (Almaån), Krycklan, Skivarp (Skivarpsån), Skövde (Ösan), Torup (Kilan), Tumba (Saxbroån), Dalbergsån, Degeå, Hässjaån, Lillån, and Lillån-Blekinge. Note that Lillån and Lillån-Blekinge are far apart (Lillån is north-west of Stockholm).

- elevation map from the Land Survey of Sweden, 50m × 50m resolution;
- terrain slope map from the Land Survey of Sweden, 50m × 50m resolution;
- soil moisture map the Swedish University of Agricultural Sciences, 2m × 2m resolution;
- land cover map the Swedish Environmental Protection Agency, 10m × 10m resolution;
- soil type map from the Geological Survey of Sweden, 10m × 10m resolution;
- soil depth map from the Geological Survey of Sweden, 10m × 10m resolution;
- hydraulic conductivity map from the Geological Survey of Sweden, 100m × 100m resolution.

The elevation map provides each coordinate's elevation above the ocean level, whereas the terrain slope map provides the slope of each coordinate (obtained by computing differences between adjacent coordinates of the elevation map).

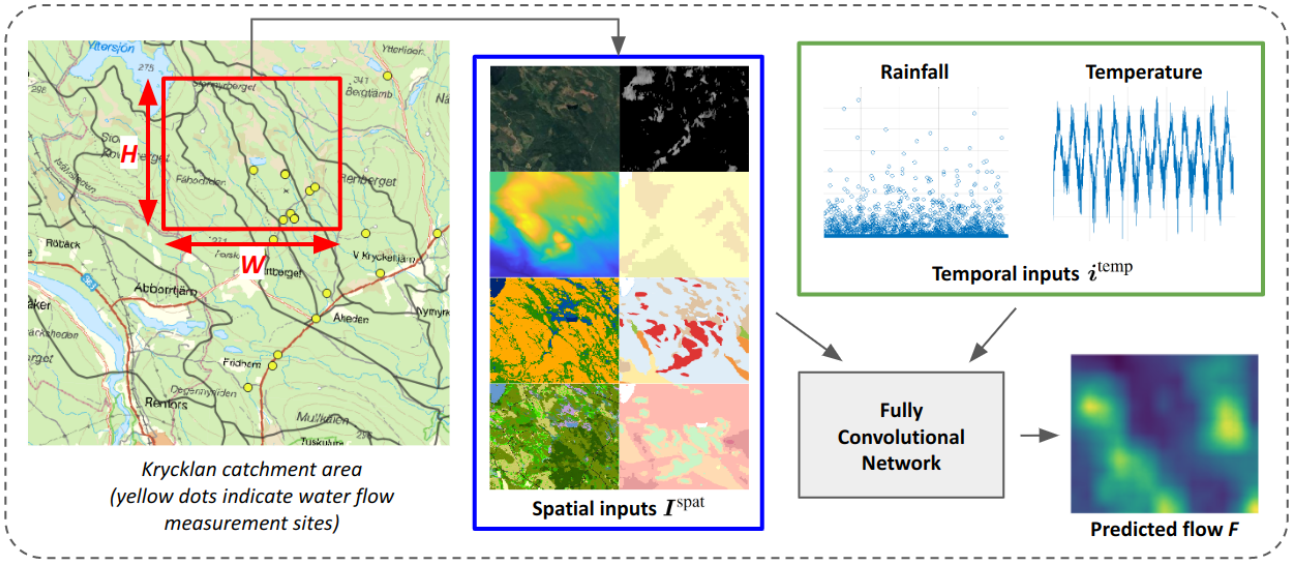

Fig. 2. Overview of our machine learning (ML) approach for dense water flow intensity prediction in catchment areas. The fully convolutional neural network receives both spatial and temporal inputs and produces a map of predicted water flow intensities. The spatial input $\boldsymbol{I}^{\text{spat}} \in \mathbb{R}^{H \times W \times C}$ represents relevant properties of the region (red rectangle, which can be located anywhere) in which to perform water flow intensity prediction (e.g. elevation map, soil moisture, land cover). The temporal inputs $\boldsymbol{i}^{\text{temp}} = \{r_j, \tau_j\}_{j=t-T}^{j=t-1}$ are the daily average rainfall amounts r_j and temperatures τ_j over the past T days, provided at times $t-T, t-T+1, \ldots, t-1$. The model then produces the flow map $\boldsymbol{F} \in \mathbb{R}^{H \times W}$, where each pixel (k, l) represents the model's estimate of the daily average water flow intensity at location (k, l) in the input map during day t. Note that our model is only trained in a very sparse set of coordinates (since water flow intensity measurement stations are very scarcely located), but the model nonetheless predicts flow intensities in every coordinate, not only in those for which contain flow intensity measurement stations.

All input data were provided as spatially aligned sets of raster maps, each of size 825×1244 pixels. For all locations except the Krycklan catchment area, there is exactly one water flow intensity measurement station. In Krycklan there are 14 such measurement stations, so in total there are 28 water flow intensity measurement stations in the dataset. In each measurement site, the daily average water flow intensity (m^3/sec) is provided. The specific length of each time series varies, but generally span several decades (in average about three decades).

In each location we also have access to two additional time series – daily cumulative rainfall (mm) and daily average temperature (degrees Celsius). These were obtained from the Swedish Meteorological and Hydrological Insitute (SMHI), and are often measured some distance away from the water flow intensity measurement sites (typically 1-3 kilometers).

It is common with missing measurements in the time series. For those time series which are used as model inputs (see Section IV), this is remedied by linearly interpolating the missing values between end points. Note that this is not done for the regressor (water flow intensity), since we only want the model to learn on and be evaluated on actual measurements.

Data preprocessing. We perform normalization of both the spatial and temporal data. Specifically, the spatial inputs are normalized to the $[0, 1]$-range by dividing with the maximum value (layer-wise) across all locations. A similar $[0, 1]$-normalization is performed for the temporal inputs (rainfall, temperature, and water flow intensity). We also tried another common normalization technique, where the variables are normalized to zero mean and unit variance, but empirically found that the $[0, 1]$-normalization works best in our setup.

IV. METHODOLOGY

In this section we provide an overview of the approach that we have developed for tackling the water flow intensity prediction task. See Figure 2 for an overview of our model and setup.

Our model leverages both spatial inputs $\boldsymbol{I}^{\text{spat}} \in \mathbb{R}^{H \times W \times C}$ and temporal inputs $\boldsymbol{i}^{\text{temp}} = \{r_j, \tau_j\}_{j=t-T}^{j=t-1}$ (cf. Section III), in order to predict water flow intensities f_t at time[1] t. The task of the model is to predict the water flow intensity $f_t^{h,w}$ at every coordinate (h, w) in a given geographical area of size $H \times W$ given $\boldsymbol{I}^{\text{spat}}$ and $\boldsymbol{i}^{\text{temp}}$. For the temporal inputs, we have chosen to only use readily available rainfall $r_{t-T}, \ldots, r_{t-1}$ and temperature data $\tau_{t-T}, \ldots, \tau_{t-1}$ for the past T days (with $T = 20$ in our setup), and not water flow intensity data $f_{t-T}, \ldots, f_{t-1}$, which is often unavailable in practice. In particular, note that water flow intensity is only measured at a very sparse subset of all coordinates in each location – and there are many locations in Sweden (and beyond) where no such measurement setups exist at all. Hence, for the model to be useful in a much larger set of contexts, it does not rely on past water flow intensities as input. However, in Section V we also compare with model variants that include past water flow intensities when predicting future flow intensities.

[1] Similar to most prior works, we target next-day prediction, but the model and approach can be extended to predict further into the future.

The spatial input $\boldsymbol{I}^{\text{spat}} \in \mathbb{R}^{H \times W \times C}$ contains relevant information regarding land and topological properties that affect the water flow intensity in any given coordinate. These spatial input layers were introduced in Section III. In our setup we let $H = W = 100$, which corresponds to a real-world area of size 1km $\times$ 1km. The number of layers C is 10 in our case (three layers for the RGB satellite images, and one layer each for the other types of spatial input).

Since the task is to predict the water flow intensity in every coordinate in a map of size $H \times W$, based (in part) on spatial inputs of size $H \times W \times C = 100 \times 100 \times 10$, we have opted for a fully convolutional neural network[2] (FCN) [11]. This architecture expects a spatial input at one end, and gives a spatial output at the other end. To achieve this, we first concatenate the spatial and temporal data $\boldsymbol{I}^{\text{spat}}$ and $\boldsymbol{i}^{\text{temp}}$ into a unified input $\boldsymbol{I} \in \mathbb{R}^{H \times W \times (C+2T)}$. The first C channels are identical to $\boldsymbol{I}^{\text{spat}}$, while the last $2T$ channels are obtained by tiling rainfall $r_{t-T}, \ldots, r_{t-1}$ and temperature data $\tau_{t-T}, \ldots, \tau_{t-1}$ into $H \times W$-dimensional maps that are concatenated along the channel-dimension (each such map contains HW copies of a single value r_i or τ_i, for $i \in \{t-T, \ldots, t-1\}$). Given an input $\boldsymbol{I}$, the water flow intensity mapping is straightforward: $\boldsymbol{F} = g_{\boldsymbol{\theta}}(\boldsymbol{I})$, where $\boldsymbol{\theta}$ denotes the learnable parameters of the FCN g. We also evaluate and compare with other architecture variants in Section V.

A. Model Training

We randomly set aside 9 of the 12 data locations for training and 3 for validating the models. Specifically, the models are evaluated in Jönköping (Tabergsån), Hässjaån and Lillån, and trained on the other 9 locations; please refer to Section III for details about the dataset. In particular, note that Lillån-Blekinge is in the training set, but it is at a vastly different location than the validation set location Lillån (cf. Figure 1).

Each training input $\boldsymbol{I}$ in a batch is generated by randomly sampling a location and time period from the training set. Once a specific location has been randomly selected, we randomly sample a sub-region of size $H \times W$ which contains a water flow measurement site at the given location – see Figure 3. This results in the spatial input $\boldsymbol{I}^{\text{spat}} \in \mathbb{R}^{H \times W \times C}$. After having sampled a spatial location, we then concatenate the temporal information $\boldsymbol{i}^{\text{temp}}$ from a randomly sampled time interval (of T consecutive days), to obtain the input $\boldsymbol{I} \in \mathbb{R}^{H \times w \times (C+2T)}$.

There are roughly $25 \cdot 100 \cdot 100 = 250,000$ different spatial training inputs (25 measurement sites, and at each site there are roughly $100 \cdot 100$ possible locations for an enclosing rectangle of size $H \times W = 100 \times 100$ – see Figure 3). Note however that there is a spatial overlap between all different rectangles at a given site, which significantly reduces the data variability, compared to if all $250,000$ different rectangles would have come from different locations. The sites have on average roughly three decades of daily rainfall, temperature and water flow averages, which means there are $30 \cdot 365 \approx 11,000$ different temporal inputs per site. Hence, in total there are roughly $250,000 \cdot 11,000 = 2.75 \cdot 10^9$ spatio-temporal training inputs. Again, however, note that there is an overlap between a large majority of these training inputs, so the training set is effectively much smaller.

[2]We use the *FCN8* model from the open-source FCN library [20].

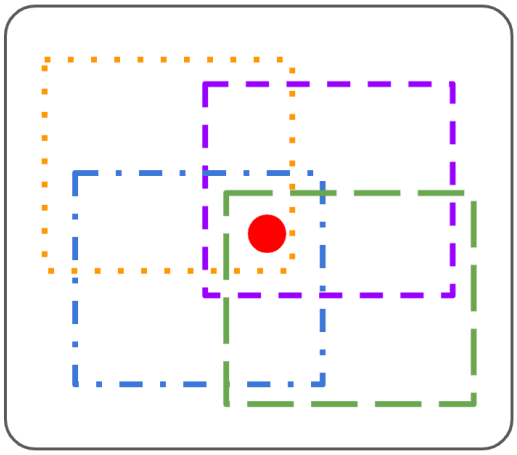

Fig. 3. Examples of possible sampled spatial locations (colored, dashed rectangles) that contain a water flow intensity measurement station (red dot). The random sampling increases the data variability (compared to e.g. always requiring the measurement station to be at the center). Since each possible spatial location has size $H \times W$, with $H = W = 100$ in our setup, the union of all possible such rectangles covers roughly 200×200 pixels, which corresponds to a real-world area of size 2km $\times$ 2km.

For model evaluation (see Section V), the end objective is to minimize the root-mean-square error (RMSE) of the predicted water flow intensities (thus note that the error is measured in m^3/sec), assessed based on ground truth flow intensities. During training, in order to balance loss smoothness with robustness to outliers, we use the Huber loss:

$$\mathcal{L}(f, f^{\text{gt}}) = \begin{cases} \frac{1}{2}\left(f - f^{\text{gt}}\right)^2 & \text{if } \|f - f^{\text{gt}}\| \leq \delta \\ \delta\left(\|f - f^{\text{gt}}\| - \frac{1}{2}\delta\right) & \text{else} \end{cases} \quad (1)$$

where f and f^{gt} denote predicted and ground truth flow intensity, respectively. We set $\delta = 1$ by default, as it is shown to result in the best performance (see Section V, where we also compare with other loss functions). The Huber loss can be seen as a combination of the commonly used MSE- and L1-losses, where the MSE-loss is applied when the error is smaller than the threshold δ, and the L1-loss is applied otherwise.

Note that for each predicted flow map $\boldsymbol{F}$, the loss (1) is only given for an extremely sparse set of coordinates (most commonly in a single point). This is because in the ground truth flow map $\boldsymbol{F}^{\text{gt}}$, we only have access to water flow measurements in very few coordinates (since the measurement sites are so sparsely located in the data). Despite this extreme loss sparsity, we show in Section V that the model generalizes well to unseen data. This finding is is in line with earlier works that have shown that it is possible to train semantic segmentation models from extremely few annotated pixels [16].

TABLE I

EXPERIMENTAL RESULTS ON THE VALIDATION SET FOR OUR MAIN MODEL, ITS ABLATED VARIANTS, AND BASELINES. WE REPORT THE ROOT-MEAN-SQUARE ERROR (RMSE; LOWER IS BETTER). COLUMN 1 REPRESENTS OUR MAIN FCN MODEL, CF. SECTION IV. COLUMNS 2-4 REPRESENT MODEL VARIANTS WHICH OMIT SOME OF THE SPATIAL INPUT LAYERS. COLUMNS 4-6 REPRESENT VARIANTS WHICH OMIT SOME TEMPORAL INPUTS. COLUMNS 7-8 REPRESENT BASELINE METHODS AGAINST WHICH TO COMPARE THE RESULTS IN COLUMNS 1-5. NOTE THAT *PREVIOUS FLOW* LEVERAGES PAST WATER FLOW INTENSITY INFORMATION THAT IS UNAVAILABLE TO THE OTHER APPROACHES.

Main model	**No-elev**	**Only-elev**	**No-soil**	**No-temp**	**No-rain**	**Half-time-hist**	**Mean-per-site**	**Previous flow**
1.35	4.74	3.22	4.49	1.91	5.91	1.40	2.05	0.59

For model parameter optimization, we use Adam [9] with batch size 64 and learning rate $2 \cdot 10^{-4}$. The model is trained for $250,000$ batches, which takes about 48 hours on the Titan V100 work station that is used for experimentation. To improve model generalization towards unseen data, we resort to the customary deep learning training technique of augmenting the data by horizontal and vertical flips of the inputs (an independent probability of 50% per flip).

V. EXPERIMENTS

In this section we present the results of our empirical model evaluations on the validation set. We first describe the various baselines and model variants in Section V-A. Then, in Section V-B, the empirical results are presented.

A. Baselines and Model Variants

We compare our main model described in Section IV against the following baselines:

- ***Mean-per-site*:** For each water flow intensity time series $\boldsymbol{f}^i = \{f^i_j\}_{j=t_1}^{j=t_{N^i}}$, where i indexes the i:th spatial location for a water flow measurement site, we return the mean $\hat{f}^i$ and use that as the predicted water flow intensity at time t (for each day t) at the i:th site. Note that this provides the optimal prediction (in terms of RMSE) in case only spatial information would be used as model input.
- ***Previous flow*:** Provides f_{t-1} as the predicted water flow intensity at time t. Note that this baseline leverages information that our model does not have access to; our model only obtains past rainfall and temperature information, not past water flow intensities. Thus *previous flow* can be regarded as a proxy for an upper bound in terms of model performance.

We also train and evaluate the following variants of our proposed ML model:

- ***No-elev*:** Omits the elevation and terrain slope maps from the set of spatial input maps.
- ***Only-elev*:** For the spatial part of the input, this model only uses the elevation and terrain slope maps. It omits the other spatial input layers.
- ***No-soil*:** Omits the soil information spatial layers (soil type, soil moisture, soil depth, land cover) from the set of spatial input maps.
- ***Half-time-history:*** Uses temporal information from the past $T = 10$ days (instead of $T = 20$ as is default).
- ***No-temp*:** Omits temperature information as an input.
- ***No-rain*:** Omits rainfall information as an input.
- ***Flow-(t-k)*:** In addition to all the spatial and temporal inputs of our main model, this model has water flow intensity information $f_{t-T-k+1}, f_{t-T-k+2}, \ldots, f_{t-k}$ as an additional temporal input when predicting the water flow intensity f_t at time t. We train and evaluate models with $k \in \{1, 2, 3\}$, i.e. models that have temporal information up to between three and one day prior to the day for which flow intensities are predicted.

Finally, we also train and evaluate the effect of variations to the main model architecture (cf. Section IV):

- ***Alt-rain-temp*:** Uses a more efficient temporal input representation, which results in the input $\boldsymbol{I}$ having dimension $H \times W \times (C+T)$ instead of $H \times W \times (C+2T)$. This is achieved by having two unique values per temporal layer (instead of only one), where every second element (spatially) is a rainfall measurement, and every second element is a temperature measurement. Note that the convolutional filters (even the first one) will have sufficiently receptive fields to observe all relevant temporal inputs in this case as well.
- ***FC-early*:** Instead of performing a concatenation of the raw temporal data along the channel dimension, this model first processes the temporal inputs through two fully connected (FC) layers, the last of which produces a $20,000$-dimensional vector. It then reshapes this vector into size $H \times W \times C^{\text{temp}} = 100 \times 100 \times 2$ and concatenates with $\boldsymbol{I}^{\text{spat}} \in \mathbb{R}^{H \times W \times C}$. This data volume of dimension $H \times W \times (C + C^{\text{temp}})$ is then run through all the layers of the FCN, as for the main model.
- ***FC-mid*:** Similar to *FC-early*, this model first processes the temporal inputs through two FC layers, but here the resulting vector has dimension $2888 = 38 \cdot 38 \cdot 2$. It then reshapes this vector into size $H^{\text{mid}} \times W^{\text{mid}} \times C^{\text{temp}} = 38 \times 38 \times 2$. Different to *FC-early*, this model does not perform the concatenation with the raw spatial data $\boldsymbol{I}^{\text{spat}} \in \mathbb{R}^{H \times W \times C}$; instead it first processes $\boldsymbol{I}^{\text{spat}}$ through the first third of the convolutional layers of the FCN. It then performs the concatenation at this stage of the FCN, followed by joint processing for the remaining two thirds of the network.

B. Empirical Results

As mentioned in Section IV-A, we randomly set aside 9 of the 12 data locations for training and 3 for validating the models. The results of our experiments on the validation set are shown in Figure 4 and Table I - IV. The evaluation metric that we report is the root-mean-square error (RMSE).

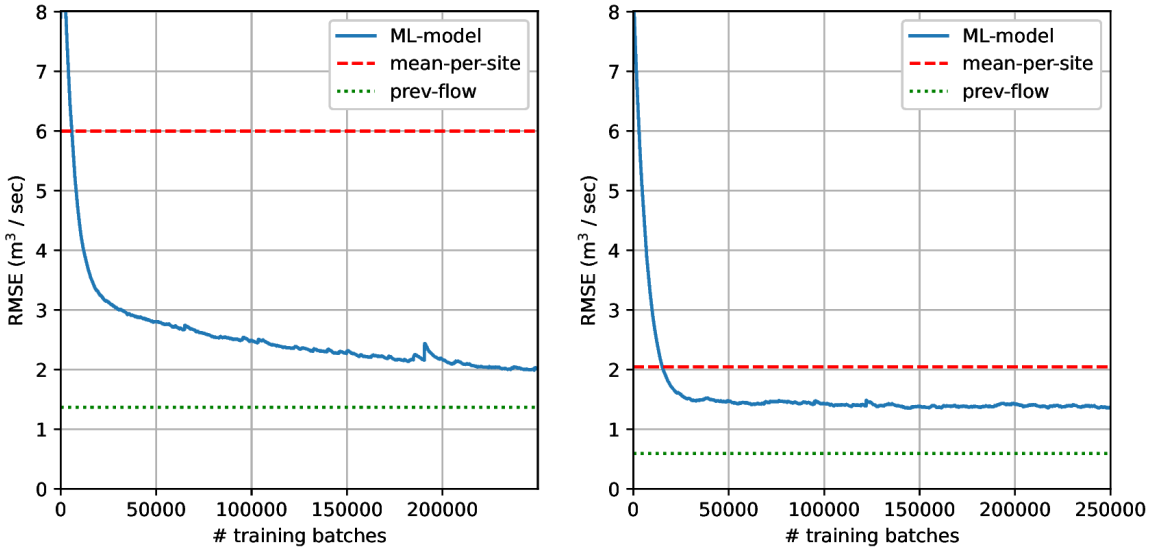

Fig. 4. Training (left) and validation (right) RMSE curves during model training for our main ML model (FCN) described in Section IV. It can be seen that the validation RMSE curve of our model flattens (marginally decreases) throughout training, i.e. the model does not begin to overfit on the training data despite the relatively small size of the training set. *Mean-per-site* and *previous flow* represent baselines, of which *previous flow* can be seen as an oracle that leverages past water flow intensity information that is unavailable to our model.

TABLE II

COMPARISON TO MODELS WHICH OBTAIN PAST WATER FLOW INTENSITIES AS INPUT. COLUMN 1 IS OUR MAIN MODEL THAT DOES NOT OBTAIN PREVIOUS FLOWS AS INPUT. PROVIDING PAST FLOW INFORMATION IMPROVES PREDICTION ACCURACY SIGNIFICANTLY, BUT NOTE THAT IN MANY CASES SUCH INFORMATION IS NOT AVAILABLE.

No flow	**Flow-(t-3)**	**Flow-(t-2)**	**Flow-(t-1)**
1.35	0.94	0.80	0.63

Due to the small size of the overall dataset (12 distinct locations), we have not yet considered a proper train-val-test split of the data. This will be done once more data of the appropriate type has been acquired. Currently however, when comparing other model variants and baselines to our main model, we report in the tables the *best* validation RMSE that was obtained during training of the respective model variant. Different to many of the alternative approaches, however, our main model's RMSE on the validation set monotonically improves throughout training (see Figure 4), and hence the results of the main model have not been 'cherry picked' at a certain optimal iteration number based on the validation set. Thus any reported improvements of the main model relative to alternatives may in fact be larger if assessed on a withheld test set.

Main results. It can be seen in Table I that our main model outperforms its ablated variants *no-elev*, *only-elev*, *no-soil*, *no-temp* and *no-rain*. In particular, the elevation and terrain slope maps are crucial, as is past rainfall information. Past temperature information is not as important, but omitting it still results in a higher error. Using rain and temperature information from the past $T = 10$ (instead of $T = 20$; see *half-time-hist*) days leads to similar results for this data.

Furthermore, our main FCN method is significantly better than the *mean-per-site* baseline, which indicates that our model has learnt to properly leverage spatio-temporal information. Our approach does however not outperform *previous flow*, which is a very strong baseline that leverages past water flow intensity information. Since such information is often hard to come by in practical scenarios, we have opted for a model that does not require past flow intensities as input, since it makes the model much more broadly applicable. Model variants which obtain previous water flow intensity information are however evaluated in Table II.

In Figure 4 we show training and validation RMSE curves during model training for our main FCN model. Note that the validation RMSE curve flattens (marginally decreases) throughout training, i.e. the model does not begin to overfit on the training data despite the relatively small size of the training set.

Effect of providing previous water flow intensity information as model input. As seen in Table II, models that receive flow information for $T = 20$ past consecutive days until three (*flow-(t-3)*), two (*flow-(t-2)*), or one (*flow-(t-1)*) day before the prediction day are significantly more accurate at predicting water flow intensity. Note however that in many practical scenarios such information is not available.

Effect of loss function. In Table III we compare the effect of using different loss functions during training; cf. (1). It is clear that the Huber loss (with $\delta = 1.0$ or $\delta = 1.1$) yields the best results, whereas the L1-loss results in the worst results.

Effect of model architecture. In Table IV we compare the effect of using different model architectures. The more

TABLE III

LOSS ANALYSIS. THE HUBER LOSS YIELDS THE LOWEST RMSE, WITH $\delta = 1.0$ AND $\delta = 1.1$ BEING BEST. THE HUBER LOSS OUTPERFORMS THE MSE LOSS, AND THE L1 LOSS YIELDS POOR RESULTS.

Huber-1.0	MSE	L1	Huber-0.8	Huber-1.1
1.35	1.55	3.17	1.47	1.35

TABLE IV

MODEL ARCHITECTURE COMPARISONS. THE *FC-EARLY* AND *FC-MID* ARCHITECTURES YIELD SIGNIFICANTLY WORSE RESULTS THAN THE MAIN MODEL AND THE *ALT-RAIN-TEMP* ARCHITECTURE.

Main model	Alt-rain-temp	FC-early	FC-mid
1.35	1.47	2.57	2.49

efficient *alt-rain-temp* architecture yields almost as good results as our main architecture, so it would be suitable to consider if compute is a limiting factor. The *FC-early* and *FC-mid* architectures yield significantly worse results.

Qualitative examples. Several qualitative examples for our main model are shown in Figure 5 - 6. Different to the quantitative experimental results above – which are only performed for the sparse set of spatial ground truth locations that are available in the data[3] – these qualitative examples shed more light into the full spatial extents of the model predictions. For example, as can be expected, higher water flow intensities are typically predicted where the terrain slope is high.

VI. DISCUSSION AND CONCLUSIONS

In this work we have introduced a fully convolutional approach for dense water flow intensity prediction in catchment areas. Our specific results were shown for Swedish basins, but the general methodology is expected to be transferable to other geographical regions.

The proposed model is able to learn and generalize from a limited training dataset. In this work, we have used training data from merely 25 measurement points (28 in total; 3 were used for evaluation). The fact that we obtain such high performance may be attributed to the training setup. In particular, the model generalization is alleviated by the fact that the model sees many slight variations of each measurement site during training, since there are many ways to select a viewpoint around a given measurement site (cf. Figure 3). To the best of our knowledge, this is the first work which models water flow intensity using a fully convolutional neural network, which allows us to provide *dense* flow predictions – in effect, we predict one flow intensity per coordinate, even though we only have annotations for 28 specific coordinates.

Since our main FCN method is significantly better than the *mean-per-site* baseline, we conclude that our model has learnt to properly leverage spatio-temporal information. Our approach does however not outperform the *previous flow* baseline, which is a very strong baseline that leverages past water flow intensity information. Such information is not available to our model, and is often hard to come by in practical scenarios. As can be seen in Table II, the FCN model variants which utilize past flow information also obtain better results. A potential avenue of future work is thus to consider our setup through a privileged learning lens, wherein flow information could be leveraged during training, but where the model must perform inference using only rainfall and temperature information (in addition to spatial information).

We hope that our work will serve as a solid stepping stone and an inspiration for further research within dense water flow modeling, which in turn could deliver useful information when it comes to future climate adaptation planning (e.g. within flood risk management) in Sweden and beyond.

[3]One can however argue that a form of (semi-)dense evaluation is being performed also in the quantitative results, since we vary the spatial coordinates of the ground truth flow intensity measurement stations in each spatial input example (cf. Figure 3). Note that the model is never aware of the spatial coordinates in which it is being evaluated.

REFERENCES

[1] Claes Bernes. En varmare värld: Växthuseffekten och klimatets förändringar-tredje upplagan, 2017.

[2] Brian Cosgrove, David Gochis, Edward P Clark, Zhengtao Cui, Aubrey L Dugger, Greg M Fall, Xia Feng, Mark A Fresch, Jonathan J Gourley, Sadiq Khan, et al. Hydrologic modeling at the national water center: Operational implementation of the wrf-hydro model to support national weather service hydrology. In *AGU Fall Meeting Abstracts*, volume 2015, pages H53A–1649, 2015.

[3] Shiheng Duan, Paul Ullrich, and Lele Shu. Using convolutional neural networks for streamflow projection in california. *Frontiers in Water*, 2:28, 2020.

[4] Jonathan M Frame, Frederik Kratzert, Daniel Klotz, Martin Gauch, Guy Shelev, Oren Gilon, Logan M Qualls, Hoshin V Gupta, and Grey S Nearing. Deep learning rainfall–runoff predictions of extreme events. *Hydrology and Earth System Sciences*, 26(13):3377–3392, 2022.

[5] Jonathan M Frame, Frederik Kratzert, Austin Raney, Mashrekur Rahman, Fernando R Salas, and Grey S Nearing. Post-processing the national water model with long short-term memory networks for streamflow predictions and model diagnostics. *JAWRA Journal of the American Water Resources Association*, 57(6):885–905, 2021.

[6] Martin Gauch, Juliane Mai, and Jimmy Lin. The proper care and feeding of camels: How limited training data affects streamflow prediction. *Environmental Modelling & Software*, 135:104926, 2021.

[7] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[8] Xiaowei Jia, Jacob Zwart, Jeffrey Sadler, Alison Appling, Samantha Oliver, Steven Markstrom, Jared Willard, Shaoming Xu, Michael Steinbach, Jordan Read, et al. Physics-guided recurrent graph model for predicting flow and temperature in river networks. In *Proceedings of the 2021 SIAM International Conference on Data Mining (SDM)*, pages 612–620. SIAM, 2021.

[9] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[10] Frederik Kratzert, Daniel Klotz, Claire Brenner, Karsten Schulz, and Mathew Herrnegger. Rainfall–runoff modelling using long short-term memory (lstm) networks. *Hydrology and Earth System Sciences*, 22(11):6005–6022, 2018.

[11] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015.

[12] Karin Mossberg Sonnek, Annika Carlsson-Kanyama, and Carl Denward. Värmens påverkan på samhället–en kunskapsöversikt för kommuner med faktablad och rekommendationer vid värmebölja. *Myndigheten för samhällsskydd och beredskap, Report No.: MSB870*, 2015.

[13] Ana Ramos Oliveira, Tiago Brito Ramos, and Ramiro Neves. Streamflow estimation in a mediterranean watershed using neural network models: A detailed description of the implementation and optimization. *Water*, 15(5):947, 2023.

[14] Tobias Salmonsson. Assessing the impacts of climate change on runoff along a climatic gradient of sweden using persist, slu masters thesis, 2013, 2014.

[15] Lisbeth Schultze, Carina Keskitalo, Irene Bohman, Robert Johannesson, Erik Kjellström, Henrik Larsson, Elisabet Lindgren, Sofie Storbjörk, and Gregor Vulturius. National expert council for climate adaptation assessment report nr 1. 2022.

[16] Gyungin Shin, Weidi Xie, and Samuel Albanie. All you need are a few pixels: semantic segmentation with pixelpick. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1687–1697, 2021.

[17] Wen-Ping Tsai, Dapeng Feng, Ming Pan, Hylke Beck, Kathryn Lawson, Yuan Yang, Jiangtao Liu, and Chaopeng Shen. From calibration to parameter learning: Harnessing the scaling effects of big data in geoscientific modeling. *Nature communications*, 12(1):5988, 2021.

[18] Song Pham Van, Hoang Minh Le, Dat Vi Thanh, Thanh Duc Dang, Ho Huu Loc, and Duong Tran Anh. Deep learning convolutional neural network in rainfall–runoff modelling. *Journal of Hydroinformatics*, 22(3):541–561, 2020.

[19] Charuleka Varadharajan, Alison P Appling, Bhavna Arora, Danielle S Christianson, Valerie C Hendrix, Vipin Kumar, Aranildo R Lima, Juliane Müller, Samantha Oliver, Mohammed Ombadi, et al. Can machine learning accelerate process understanding and decision-relevant predictions of river water quality? *Hydrological Processes*, 36(4):e14565, 2022.

[20] Ketaro Wada. pytorch-fcn: PyTorch Implementation of Fully Convolutional Networks. `https://github.com/wkentaro/pytorch-fcn`, 2017.

[21] Andreas Wunsch, Tanja Liesch, and Stefan Broda. Groundwater level forecasting with artificial neural networks: a comparison of long short-term memory (LSTM), convolutional neural networks (CNNs), and non-linear autoregressive networks with exogenous input (NARX). *Hydrology and Earth System Sciences*, 25(3):1671–1687, 2021.

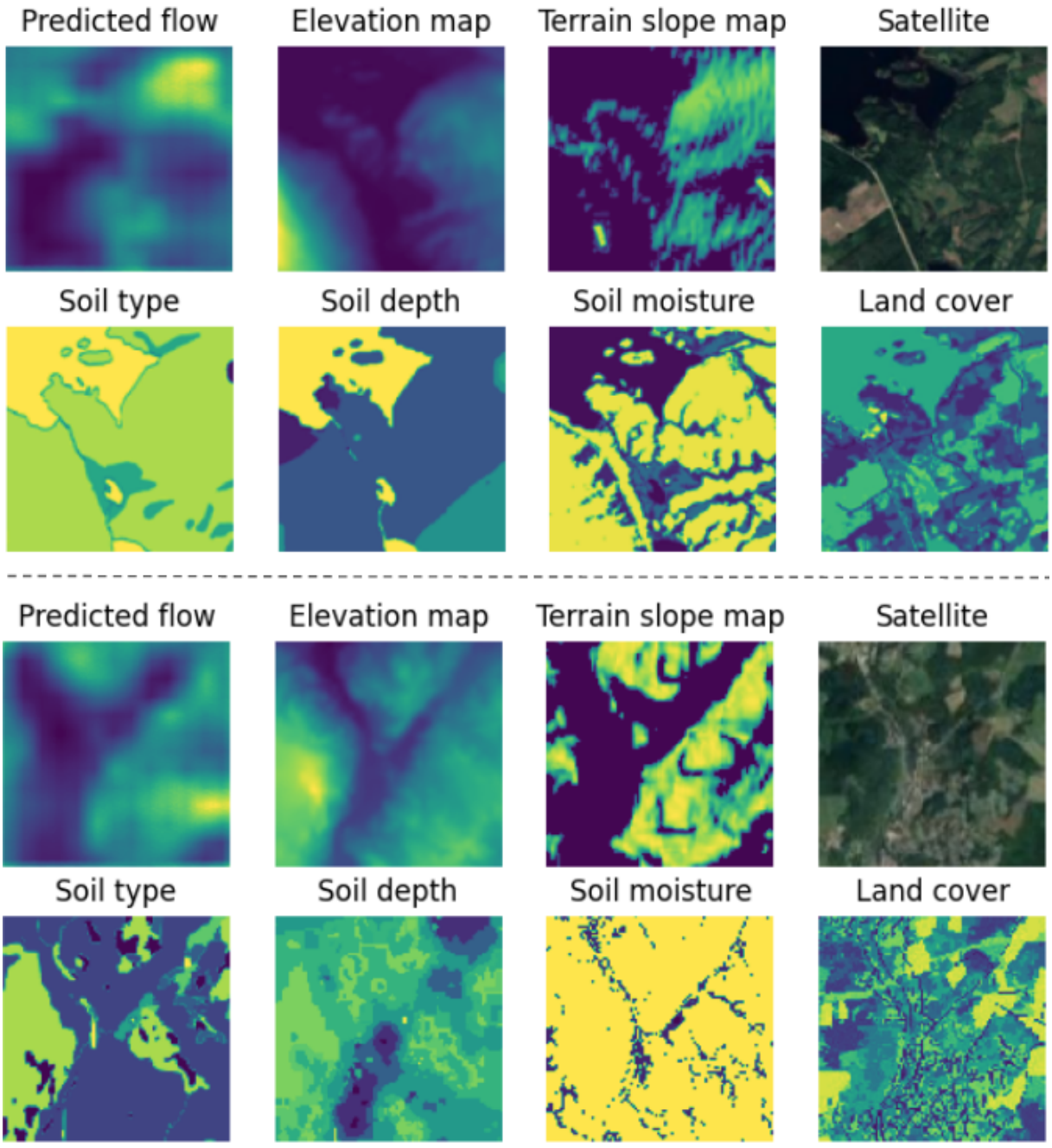

Fig. 5. Two qualitative examples for our main model on the validation set (examples differentiated by the dashed horizontal line). In each example, the top-left image represents the predicted water flow intensities at the given area; darker blue means lower intensity, while brighter yellow means higher intensity. The other seven images represent various spatial input layers to the FCN model. For all images except the satellite image, the maximum color intensity is individually normalized so that variations within images become as visible as possible. As can be expected, in both examples, higher flow intensities are typically predicted where the terrain slope is higher. In the example above the dashed line, the model also predicts relatively high flow intensities on the lake that can be observed at the top-left of the satellite image. Note however that the training set contains no ground truth water flow intensities on lakes, and thus the model has never been able to adapt to what is reasonable in terms of flow intensity on lakes.

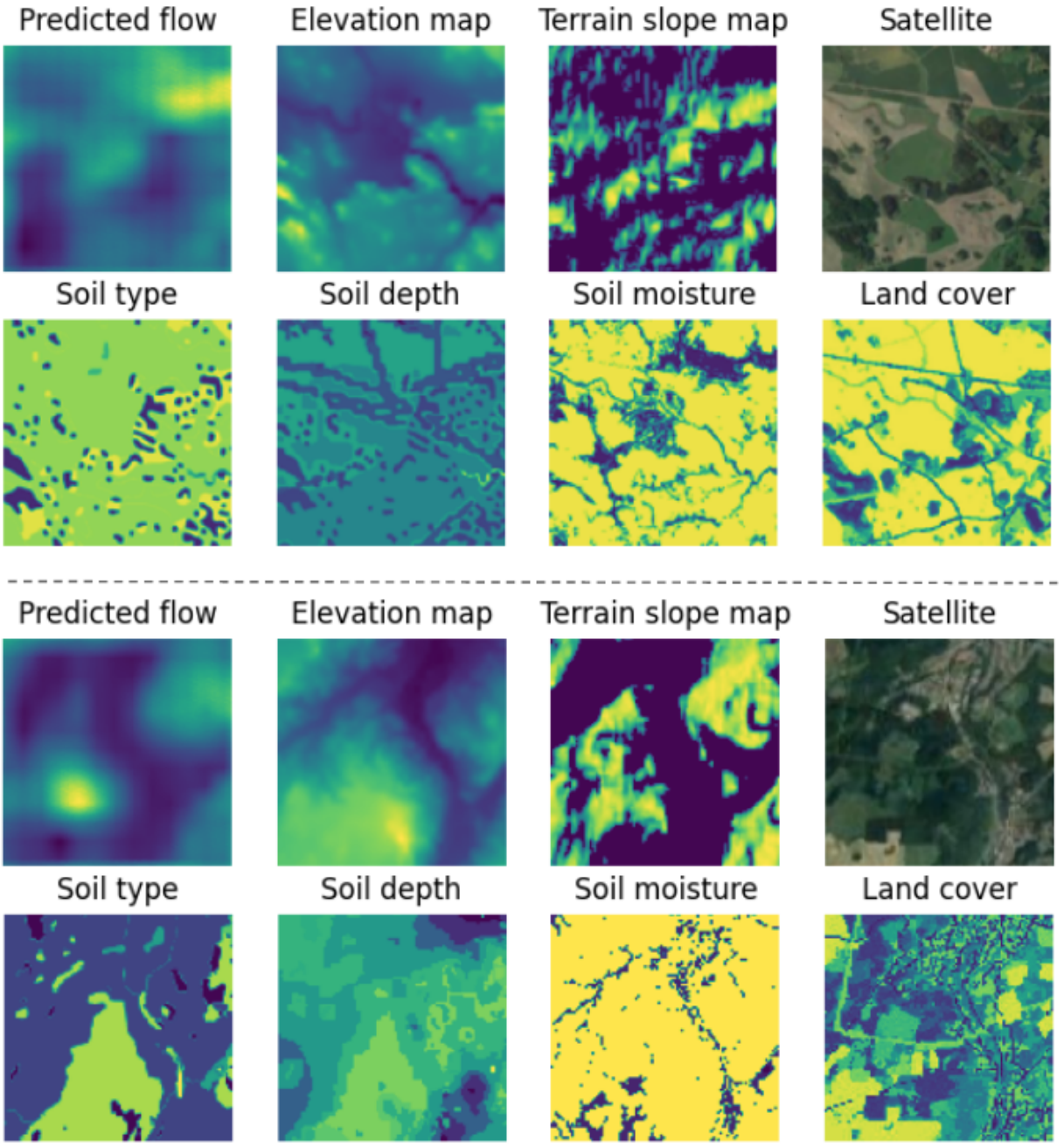

Fig. 6. Two additional qualitative examples for our main model on the validation set (examples differentiated by the dashed horizontal line). In the example above the dashed line, the predicted flow is moderately high within most of the map. A peak in terms of predicted flow can be seen at a corresponding peak within the terrain slope map. In the example below the dashed line, in can again be seen that the predicted flow is typically relatively higher where the terrain slope map is higher

Aerial View Localization with Reinforcement Learning: *Towards Emulating Search-and-Rescue*

Aleksis Pirinen[1], Anton Samuelsson[2], John Backsund[2] and Kalle Åström[2]
[1]RISE Research Institutes of Sweden
[2]Centre for Mathematical Sciences, Lund University, Sweden
{aleksis.pirinen@ri.se, anton.b.samuelsson@gmail.com,
j.backsund@gmail.com, karl.astrom@math.lth.se}

***Abstract*— Climate-induced disasters are and will continue to be on the rise, and thus search-and-rescue (SAR) operations, where the task is to localize and assist one or several people who are missing, become increasingly relevant. In many cases the rough location may be known and a UAV can be deployed to explore a given, confined area to precisely localize the missing people. Due to time and battery constraints it is often critical that localization is performed as efficiently as possible. In this work we approach this type of problem by abstracting it as an *aerial view goal localization* task in a framework that emulates a SAR-like setup without requiring access to actual UAVs. In this framework, an agent operates on top of an aerial image (proxy for a search area) and is tasked with localizing a goal that is described in terms of visual cues. To further mimic the situation on an actual UAV, the agent is not able to observe the search area in its entirety, not even at low resolution, and thus it has to operate solely based on partial glimpses when navigating towards the goal. To tackle this task, we propose *AiRLoc*, a reinforcement learning (RL)-based model that decouples exploration (searching for distant goals) and exploitation (localizing nearby goals). Extensive evaluations show that AiRLoc outperforms heuristic search methods as well as alternative learnable approaches, and that it generalizes across datasets, e.g. to disaster-hit areas without seeing a single disaster scenario during training. We also conduct a proof-of-concept study which indicates that the learnable methods outperform humans on average. Code and models have been made publicly available at `https://github.com/aleksispi/airloc`.**

I. INTRODUCTION

Recent technological developments of unmanned aerial vehicles (UAVs) and satellites have resulted in an enormous increase in the amount of aerial view landscape and urban data that is available to the public [4], [15], [28], [12], [38], [26], [36]. An important application area of UAVs is within search-and-rescue (SAR) operations, where the task is to localize and assist one or several people who are missing, for example after a natural disaster. It may often be the case that the people in need are known to be within a confined area, such as within a specific neighborhood or city block. In such a scenario, a UAV can be used to explore the area from an aerial perspective to precisely localize and subsequently assist the missing people. Obviously, controlling the UAV in an informed and intelligent manner, rather than exhaustively scanning the whole area, could significantly improve the likelihood of succeeding with the operation.

In this paper, we propose a novel setup and task formulation that allows for controllable and reproducible development of and experimentation with systems for UAV-based SAR operations.[1] More specifically, we abstract the problem within a framework that emulates a SAR-like setup without requiring access to actual UAVs. In this framework, an agent operates on top of an aerial image (proxy for a specific search area) and is tasked with localizing a goal for which coordinates are not available, but where some visual cues of the goal are provided. For our task, which we denote *aerial view goal localization*, we assume that the visual cues are given in terms of a top-view observation of the goal within the search area (see Fig. 1). This provides a streamlined proxy setup, but note that in a real SAR operation such cues could instead be provided e.g. by the missing people, assuming they have been able to send information about their surroundings (e.g. ground-level images). The active localization methodologies we propose can easily be extended to allow for more flexible goal specifications, for example by integrating an off-the-shelf geo-localization module.

There are many cases where GPS coordinates of the goal location are not available, or where such information is not reliable (e.g. because global satellite navigation systems are susceptible to radio frequency interruptions and fake signals). Hence there is a need for robust aerial localization systems that do not rely on global positional information, but that can operate reliably based on visual information alone. Moreover, to further mimic the situation on an actual UAV, it is assumed in our task that only a partial glimpse of the search area can be observed at the same time. In many cases, a UAV could elevate to a higher altitude to get a generic (lower-resolution) sense of the whole search area, but there are also conditions which makes this impractical, e.g. if the battery of the UAV is running low. Adverse weather conditions could also make it risky or impossible to operate at a high altitude.

To tackle our suggested aerial view goal localization task, we propose *AiRLoc*, a reinforcement learning (RL)-based model that decouples exploration (searching for distant goals) and exploitation (localizing nearby goals) – see Fig. 1. Extensive experimental results show that AiRLoc outperforms heuristic search methods and alternative learnable approaches. The results also show that AiRLoc generalizes across datasets, e.g. to disaster-hit areas without seeing a

[1]Also relevant for many types of environmental monitoring applications, e.g. in forestry management.

single disaster scenario during its training phase. We also conduct a proof-of-concept study which indicates that this task is difficult even for humans.

II. RELATED WORK

Several prior works have proposed methods for autonomous control a UAVs [27], [13], [6], [3], [24], [41], [19]. Many of these works (e.g. [27], [24], [41]) revolve around methodologies for efficient scanning of large areas (e.g. agricultural landscapes) such that certain types of global-level downstream inferences – such as determining the health status of a field of crops – can be accurately performed based on a limited number of high-resolution observations. Aside from differing in task formulation (ours requiring precise localization of a particular goal, while the aforementioned works often revolve around global-level inference), these prior works assume access to a global lower-resolution observation of the whole area of interest, while we do not. There are also works that are closer to us in terms of task setup [3], [13], [6]. For example, [3] propose a hierarchical planning approach for a goal reaching task, where a rough plan is first proposed using A*. This rough plan is subsequently used as an initial guess by a finer-grained planner which parametrizes the initial trajectory as continuous B-splines and performs trajectory optimization. Different from us, their system assumes access to ground truth detections of moving objects and ground classifications.

Our work is also related but orthogonal to the increasingly studied problem of geo-localization [35], [30], [42], [40], [20], [33]. Such works aim to infer relationships between two or more images from different perspectives, e.g. predicting the satellite or drone view corresponding to a ground-level image. Most such methods perform this task by an exhaustive comparison within a large image set, and are thus very different to our setup which instead revolves around minimizing the amount of observations when performing localization. However, our proposed methodologies could further benefit from incorporating geo-localization methods. For example, if the goal location is specified from a ground-level perspective, which may be more realistic in practice, geo-localization methods can be used to match the top-view images observed by our proposed method during goal localization.

From a pure task formulation perspective, and setting aside the application areas, our setup may be most closely related to embodied image goal navigation [1], [43], [14]. In this framework, an agent is tasked to navigate in a first-person perspective within a 3d environment towards a goal location which is specified as an image within the environment. On the one hand, the embodied setting may sometimes be more challenging than our setup, since the exploration trajectories are typically longer (as the agent moves a significantly smaller extent per action) and because exploration is performed among obstacles (e.g. walls and furniture). On the other hand, embodied first person agents may often observe the goal from far away (e.g. from the other side of a newly entered room), while our formulation is more challenging in that the goal can never be observed in any way prior to reaching it.

To the best of our knowledge, in addition to us relatively few prior works have considered inference based solely on partial glimpses of an underlying image [21], [22]. In contrast, most earlier RL-based methods that have been proposed for computer vision tasks – e.g. for object detection [5], [8], [18] and aerial view processing [29], [2] – assume access to at least a low-resolution version of the entire scene or image being processed. Even the seminal work by [16] uses lower-resolution full image input in addition to high-resolution partial glimpses during its sequential processing, even though in principle it may be possible to re-design the system to operate based on high-resolution glimpses alone.

III. AERIAL VIEW GOAL LOCALIZATION

In this section we first explain in detail our proposed aerial view goal localization task and framework (§III-A). Then, in §III-B, we explain AiRLoc, our reinforcement learning (RL)-based approach for tackling this task. See Fig. 1 for an overview. Finally, §III-C describes the baseline methods we have developed and that we evaluate and compare with AiRLoc in §IV.

A. Task Description

The task is executed by an agent within a *search area*, which is discretized as an $M \times N$ grid that is layered on top of a given aerial image (with a small distance between each grid cell, to avoid overfitting models to edge artefacts). Every grid cell within the search area corresponds to a valid position $\boldsymbol{p}_t$ of the agent, and the agent can only directly observe the image content $\boldsymbol{O}_t$ of its current cell. In each episode, one of the grid cells corresponds to the goal that the agent should localize. The image content of the goal cell is denoted $\boldsymbol{O}^{\text{goal}}$ and its position is denoted $\boldsymbol{p}^{\text{goal}}$. Note that the goal position $\boldsymbol{p}^{\text{goal}}$ is *never* observed by the agent; it is only used to determine if the agent is successful. The task is considered successfully completed as soon as the agent's current position $\boldsymbol{p}_t$ and the goal position $\boldsymbol{p}^{\text{goal}}$ coincide,[2] i.e. when $\boldsymbol{p}_t = \boldsymbol{p}^{\text{goal}}$.

In each episode, the agent's start location $\boldsymbol{p}_0$ and the goal location $\boldsymbol{p}^{\text{goal}}$ are sampled at uniform random within the search area ($\boldsymbol{p}_0 \neq \boldsymbol{p}^{\text{goal}}$). The agent then moves around until it either reaches the goal ($\boldsymbol{p}_t = \boldsymbol{p}^{\text{goal}}$), or a maximum number of steps T have been taken. This limit T is included to represent time and resource constraints. In our task formulation, an agent has eight possible actions, which correspond to moving to any of its eight adjacent locations (grid cells). An agent may in general move outside the search area, and if so, the agent receives an entirely black observation. There is never any advantage to moving outside the search area, and thus it should be avoided (it is easy to avoid given $\boldsymbol{p}_t$).

[2] A reasonable next step would be to require that an agent has to declare when it has reached its goal.

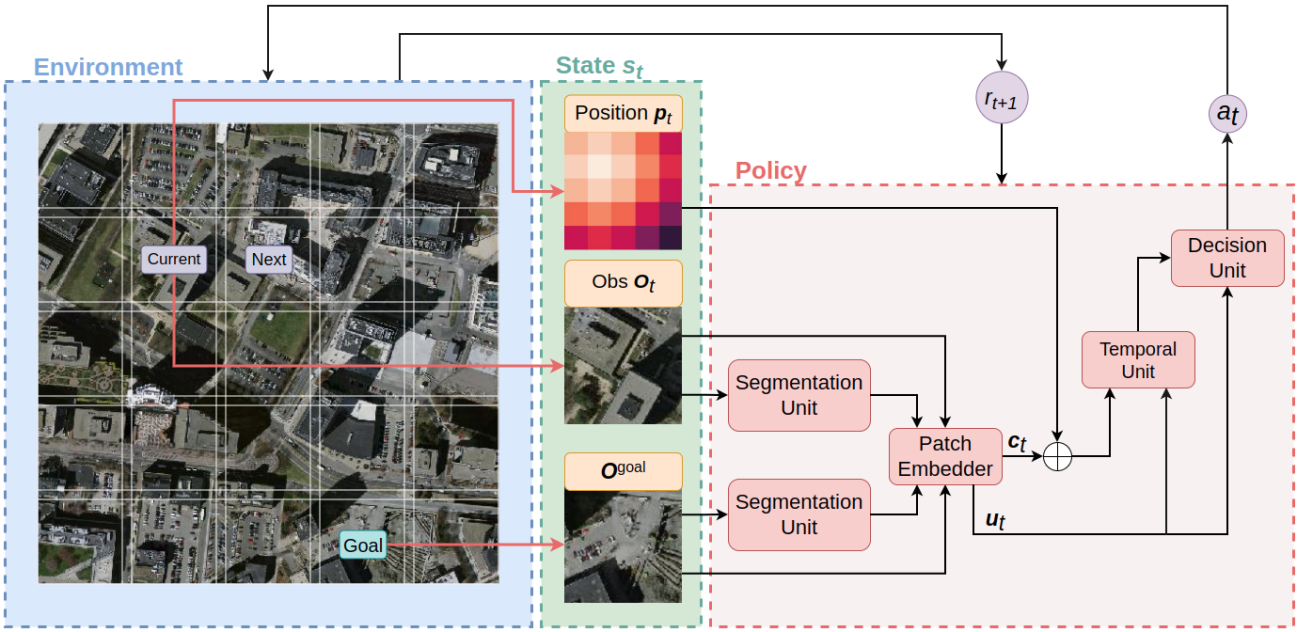

Fig. 1. Overview of *AiRLoc*, our RL-based agent for aerial view goal localization. The state s_t consists of the agent's current position $\boldsymbol{p}_t$, its currently observed patch $\boldsymbol{O}_t$, and the goal patch $\boldsymbol{O}^{\text{goal}}$. First, segmentation masks for $\boldsymbol{O}_t$ and $\boldsymbol{O}^{\text{goal}}$ are computed, and $\boldsymbol{O}_t$, $\boldsymbol{O}^{\text{goal}}$ and their segmentations are then fed through a patch embedder to generate a common representation $\boldsymbol{c}_t$. The positional encoding $\boldsymbol{p}_t$ is then added to $\boldsymbol{c}_t$, and the sum, together with an exploitation prior $\boldsymbol{u}_t$ (see §III-B), are subsequently processed by an LSTM, whose output is fed to a decision unit. The decision unit also receives $\boldsymbol{u}_t$ and outputs an action probability distribution $\pi(\cdot|s_t)$. A movement a_t is then sampled from $\pi(\cdot|s_t)$, which results in the next state s_{t+1} and reward r_{t+1} (a reward is provided during training only). The process is repeated, either until the agent reaches the goal, or until a maximum number of steps T have been taken. Note that AiRLoc never observes the full search area, not even at a low resolution.

B. AiRLoc Model

In this section we describe *AiRLoc*, the reinforcement learning (RL)-based model we propose for tackling the aerial view goal localization task; see an overview in Fig. 1.

States, actions and rewards. The state s_t contains the currently observed patch $\boldsymbol{O}_t$, the goal patch $\boldsymbol{O}^{\text{goal}}$, and an encoding $\boldsymbol{p}_t \in \mathbb{R}^{256}$ of the agent's position. As described above, AiRLoc has eight possible actions a_t, which correspond to moving to any of its adjacent locations. During training, a negative reward is provided for each action that does not move the agent into the goal location, and a positive reward is provided when the goal is found. Specifically, after taking action a_{t-1} in state s_{t-1} the reward $r_t = 3 \cdot \mathbb{1}\left(\boldsymbol{p}_t = \boldsymbol{p}^{\text{goal}}\right) - 1$ is provided, where $\mathbb{1}$ is the indicator function.

Policy overview In each step, the state s_t is processed by four modules to generate the current action distribution $\pi_{\boldsymbol{\theta}}(*|s_t)$, where $\boldsymbol{\theta}$ denotes all learnable parameters. First, $\boldsymbol{O}_t$ and $\boldsymbol{O}^{\text{goal}}$ are passed through a pretrained *segmentation unit* (a U-net [23]) which predicts building segmentation masks for $\boldsymbol{O}_t$ and $\boldsymbol{O}^{\text{goal}}$, respectively. Second, $\boldsymbol{O}_t$ and $\boldsymbol{O}^{\text{goal}}$ and their segmentations are passed through a *patch embedder* which yields a low-dimensional embedding $\boldsymbol{c}_t \in \mathbb{R}^{256}$ of what the agent observes and what it aims to localize. The patch embedder also outputs an exploitation prior $\boldsymbol{u}_t \in \mathbb{R}^8$ (described more below). Third, $\boldsymbol{p}_t$ is added to $\boldsymbol{c}_t$ and the result and $\boldsymbol{u}_t$ are passed to an LSTM-based *temporal unit* [10] which integrates information over time. Finally, the LSTM output and $\boldsymbol{u}_t$ are passed to a *decision unit* which yields the probability distribution $\pi_{\boldsymbol{\theta}}(*|s_t)$. This decision unit first projects the LSTM's output into the action space dimensionality, then adds the exploitation prior $\boldsymbol{u}_t$, and finally generates an action distribution using softmax. Note that we use an LSTM rather than a Transformer for the temporal unit, since we want to keep the overall architecture lightweight – the model weights occupy less than 4 MB of memory, and inference can be efficiently performed even without a GPU.

Patch embedder. The patch embedder should extract relevant information about the relationship between $\boldsymbol{O}_t$ and $\boldsymbol{O}^{\text{goal}}$. To achieve this, we use an architecture similar to that by [7], who consider a self-supervised visual representation learning task where the spatial displacement between a pair of adjacent random crops from an image should be predicted. Note that when the start location $\boldsymbol{p}_0$ is adjacent to the goal location $\boldsymbol{p}^{\text{goal}}$, and when the movement budget $T = 1$, our task becomes equivalent to the representation learning task introduced by [7]. Our patch embedder architecture consists of two parallel branches with four convolutional layers (ReLUs and max pooling are applied between layers). First, $\boldsymbol{O}_t$ and $\boldsymbol{O}^{\text{goal}}$, with their segmentations channel-wise concatenated, are fed separately into one branch each. To enable early information sharing between the agent's current patch and the goal patch, after two convolutional layers, the outputs of the two branches

are concatenated and sent through the rest of their respective branches. The two resulting 128-dimensional embeddings are then concatenated and the result is passed through a dense layer with output $\boldsymbol{c}_t \in \mathbb{R}^{256}$.

Pretraining backbone vision components is common in RL setups, since it often yields a higher end performance [25], [17], [32], [37], [39]. We therefore pretrain the patch embedder in the same self-supervised fashion as [7]. During pretraining, another dense layer (with input $\boldsymbol{c}_t$) is attached to produce an 8-dimensional output $\boldsymbol{u}_t$ which is fed to a softmax function. The eight outputs correspond to the possible locations of $\boldsymbol{O}^{\text{goal}}$ relative to $\boldsymbol{O}_t$, assuming these are adjacent. When using the patch embedder within AiRLoc, we take advantage of both $\boldsymbol{c}_t$ and $\boldsymbol{u}_t$, cf. Fig. 1. Note that $\boldsymbol{u}_t$ can be interpreted as an *exploitation prior*, as it is specifically tuned towards localizing ('exploiting') adjacent goals. Thus, feeding $\boldsymbol{u}_t$ to the temporal unit as well as directly to the decision unit allows AiRLoc to learn when to explore and when to exploit (without $\boldsymbol{u}_t$, the same policy must be able to both localize adjacent goals *and* explore far-away goals). The choice of using both $\boldsymbol{c}_t$ and $\boldsymbol{u}_t$ is empirically justified in §IV-B.

Positional encoding. Positional information is represented similarly to Transformers [31]. Note that AiRLoc never receives global positional information, i.e. it is always relative to a given search area. Such information may be available during SAR within a confined area, where a UAV can keep track of its location relative to the borders of this area. Let (x, y) denote the agent's coordinates within the $M \times N$-sized search area (thus $x \in \{0, \ldots, M-1\}$, $y \in \{0, \ldots, N-1\}$). Then the i:th element p_t^i of the positional encoding vector $\boldsymbol{p}_t \in \mathbb{R}^d$ (with d even; for us $d = 256$) is given by:

$$p_t^i = \begin{cases} \cos\left(x/100^{2(i-1)/(d/2)}\right) & \text{if } i \in \{1, 3, \ldots, d/2-1\} \\ \sin\left(x/100^{2i/(d/2)}\right) & \text{if } i \in \{2, 4, \ldots, d/2\} \\ \cos\left(y/100^{2(i-1)/(d/2)}\right) & \text{if } i \in \{d/2+1, \ldots, d-1\} \\ \sin\left(y/100^{2i/(d/2)}\right) & \text{if } i \in \{d/2+2, \ldots, d\} \end{cases} \tag{1}$$

Policy training. To learn the parameters of AiRLoc, we first pretrain the patch embedder in a self-supervised fashion (without RL) as described above. We then freeze the patch embedder weights and train the rest of AiRLoc using REINFORCE [34]. We employ within-batch reward normalization based on distance left to the goal, i.e. rewards associated with states of equal distance to the goal are grouped and normalized to zero mean and unit variance. We use a pretrained segmentation unit (one can simply use an off-the-shelf aerial view segmentation model) and it is not refined during policy training.

C. Baselines

In §IV we compare AiRLoc with the following baselines:

- ***Priv random*** selects actions randomly, with two exceptions: i) it cannot move outside the search area; ii) it avoids previous locations.
- ***Local*** selects actions by repeatedly calling the pretrained patch embedder (which assumes the goal is adjacent to the current location).
- ***Priv local*** is the same as *Local* but with the privileged movement restrictions of *Priv random*.
- ***Human*** represents the average human performance from a proof-of-concept evaluation with 19 subjects (see details in the appendix).

IV. Experiments

In this section we extensively evaluate and compare AiRLoc and the various baselines described in §III-B and §III-C, respectively. First we however describe what datasets and evaluation metrics we use, explain different variants of AiRLoc, and provide some further implementation details.

Datasets. We mainly use *Massachusetts Buildings* (*Masa*) by [15] for development and evaluation (70% for training; 15% each for validation and testing). The data contains images of Boston and the surrounding suburban and forested areas. It depicts houses, roads and other clearly identifiable man-made structures, but also woods and less developed regions. The data also includes segmentation masks for buildings, which are used to separately train the segmentation unit (cf. Fig. 1) that is used by most of the learnable models in the results below. Models are also evaluated on the *Dubai* dataset [28], which also depicts urban regions, although the surrounding areas are instead dry deserts. This dataset is hence used to assess the generalization of the various methods. Finally, we also train and evaluate on the *xBD* dataset by [9], which contains satellite images from various regions both before (*xBD-pre*) and after (*xBD-disaster*) various natural distastes, e.g. wildfires and floods. In this case the models are trained on non-disaster-hit data from *xBD-pre* and evaluated on *xBD-disaster*, where we also ensure that the training data depicts other geographical areas than those in *xBD-disaster*.

Evaluation metrics. We use the following evaluation metrics. ***Success*** is the percentage of episodes where the goal is reached. ***Steps*** is the average number of actions taken per episode (for failure episodes this is set to the movement budget T). ***Step ratio*** measures the average ratio between the taken number of steps and the minimum number of steps required (lower is better). It is only computed for successful trajectories. ***Residual distance*** measures the average distance between the final location relative to the goal location in unsuccessful episodes (lower is better). Finally, ***Runtime*** is the average runtime per episode.

AiRLoc variants. We also train and evaluate several ablated variants of AiRLoc. ***No sem seg*** omits the segmentation unit and uses only RGB patches in the patch embedder (which is instead pretrained with RGB-only inputs). ***No residual*** omits $\boldsymbol{u}_t$ in the decision unit, but not in the temporal unit, cf. Fig. 1. Finally, ***no prior*** entirely discards the prior $\boldsymbol{u}_t$ in the architecture.

TABLE I

RESULTS ON THE TEST SET OF *Massachusetts Buildings* (MOVEMENT BUDGET $T = 10$ AND $T = 14$ FOR SETUPS OF SIZES 5×5 AND 7×7, RESPECTIVELY). FOR BOTH SEARCH AREA SIZES, THE SUCCESS RATE OF AIRLOC IS HIGHER THAN FOR THE BASELINES. MID-LEVEL VISION CAPABILITIES (SEMANTIC SEGMENTATION) ARE CRUCIAL FOR AIRLOC'S PERFORMANCE. THE STANDARD LOCAL APPROACH PERFORMS POORLY AND IS SIGNIFICANTLY IMPROVED BY IMPOSING THE PRIVILEGED MOVEMENT CONSTRAINTS. THE TIME PER EPISODE IS LOW FOR ALL METHODS.

Agent type	Success	Step ratio	Steps	Residual distance	Runtime
AiRLoc (5x5)	67.6 %	1.45	6.2	2.4	120 ms
Priv local (5x5)	64.2 %	1.59	6.5	2.4	117 ms
Local (5x5)	24.7 %	1.47	8.1	7.0	138 ms
Priv random (5x5)	41.0 %	2.56	8.0	1.6	48 ms
AiRLoc (7x7)	59.0 %	1.52	9.4	3.3	188 ms
Priv local (7x7)	56.3 %	1.72	9.9	3.4	178 ms
Local (7x7)	17.8 %	1.20	11.9	8.7	202 ms
Priv random (7x7)	25.2 %	1.82	12.3	3.5	74 ms
AiRLoc (no sem seg, 5x5)	61.7 %	1.54	6.7	2.4	94 ms
Priv local (no sem seg, 5x5)	61.6 %	1.67	6.8	2.4	88 ms
Local (no sem seg, 5x5)	20.5 %	1.28	8.4	6.2	92 ms
AiRLoc (no sem seg, 7x7)	52.5 %	1.61	10.1	3.5	141 ms
Priv local (no sem seg, 7x7)	51.1 %	1.89	10.2	3.3	133 ms
Local (no sem seg, 7x7)	14.1 %	1.37	12.4	8.0	136 ms

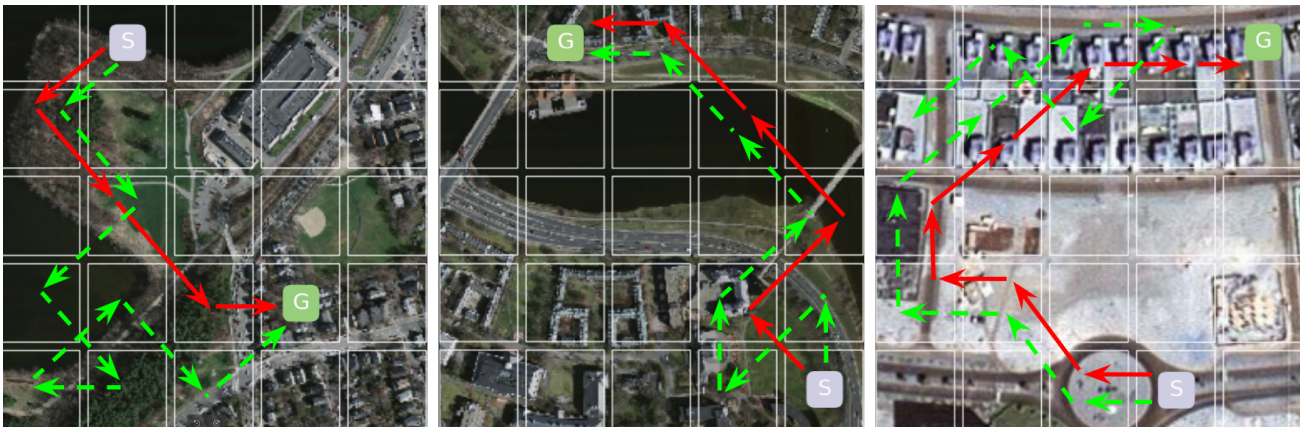

Fig. 2. Examples of AiRLoc (red) and *Priv local* (dashed green) on the test set of *Masa* (left, middle) and *Dubai* (right). Left: AiRLoc takes the same first two actions as *Priv local* and then takes the shortest path to the goal ('G'). *Priv local* also reaches the goal. Middle: AiRLoc first deviates from *Priv local* and then follows the same path. AiRLoc reaches the goal faster. Right: AiRLoc follows the same path as *Priv local* until it is adjacent to the goal and then moves into the goal, while *Priv local* fails.

Implementation details. All methods are implemented in, trained and evaluated using PyTorch. Training AiRLoctakes 30h on a Titan V100 GPU. To learn the parameters of the policy networks, we use REINFORCE [34] with Adam [11], batch size 64, search area size $M \times N = 5 \times 5$, movement budget $T = 10$, learning rate 10^{-4}, and discount $\gamma = 0.9$. The grid cells of the search areas are of size $48 \times 48 \times 3$, with 4 pixels between each other to avoid overfitting models to edge artefacts (each cell corresponds to roughly 100×100 meters). Each model is trained until convergence on the validation set (typically happens within 50k batches). We apply left-right and top-down flipping of images (search areas) as data augmentation. The AiRLoc variants are trained with five random network initializations each, and the results for the median-performing models on the validation set are reported below. AiRLoc is not seed sensitive, as shown in §IV-C. Unless otherwise specified, all models are evaluated in deterministic mode, i.e. the most probable action is selected in each step. All models are evaluated on the exact same start configurations for fair comparisons.

A. Main Results

In Table I we compare AiRLoc to the heuristic random and learnable local baselines on the test set of *Massachusetts Buildings* (*Masa*). AiRLoc obtains a higher success rate than the baselines, both in search areas of size 5×5 and 7×7 (AiRLoc is only trained in the 5×5 setting). AiRLoc and *Priv local* have roughly the same runtime per trajectory, and note that all methods have runtimes that would be negligible compared to the movement overhead of an actual UAV. It is also clear that the segmentation model is crucial, which is in line with prior works that find that mid-level vision capabilities are important for high performance in

TABLE II

AIRLOC AND BASELINES EVALUATED ON PREVIOUSLY UNSEEN *Dubai* DATA (MOVEMENT BUDGET $T = 10$ AND $T = 14$ FOR SETUPS OF SIZES 5×5 AND 7×7, RESPECTIVELY). AIRLOC AND THE PRIVILEGED LOCAL APPROACH GENERALIZE VERY WELL TO THIS OUT-OF-DOMAIN DATA. NOTE THAT AIRLOC IS THE MOST SUCCESSFUL METHOD IN ALL SETTINGS, OFTEN BY A LARGE MARGIN.

Agent type	Success	Step ratio	Steps	Residual distance	Runtime
AiRLoc (5x5)	68.8 %	1.52	6.3	2.4	126 ms
Priv local (5x5)	65.6 %	1.59	6.5	2.4	113 ms
Local (5x5)	23.5 %	1.23	8.2	6.6	136 ms
Priv random (5x5)	41.0 %	1.96	8.0	2.5	48 ms
AiRLoc (7x7)	57.2 %	1.54	9.7	3.4	194 ms
Priv local (7x7)	53.7 %	1.85	10.2	3.6	184 ms
Local (7x7)	15.5 %	1.25	12.2	7.9	207 ms
Priv random (7x7)	26.9 %	1.64	12.0	3.5	72 ms
AiRLoc (no sem seg, 5x5)	67.1 %	1.59	6.5	2.4	91 ms
Priv local (no sem seg, 5x5)	65.1 %	1.67	6.6	2.5	86 ms
Local (no sem seg, 5x5)	23.3 %	1.25	8.2	6.6	90 ms
AiRLoc (no sem seg, 7x7)	48.6 %	1.56	10.3	3.3	144 ms
Priv local (no sem seg, 7x7)	41.9 %	1.69	10.8	3.4	140 ms
Local (no sem seg, 7x7)	15.0 %	1.28	12.3	7.6	135 ms

TABLE III

RESULTS ON SCENARIOS DEPICTING VARIOUS NATURAL DISASTERS (*xBD-disaster*) FOR MODELS TRAINED IN TWO DIFFERENT WAYS. COLUMNS 1 - 3: AIRLOC GENERALIZES QUITE WELL FROM HAVING BEEN TRAINED ON AN ENTIRELY DIFFERENT DATASET (*Masa*), WHICH CONTAINS SATELLITE IMAGES OF NON-DISASTER-HIT URBAN AREAS, TO DISASTER-HIT AREAS AT VARIOUS OTHER SPATIAL LOCATIONS. COLUMNS 4 - 6: RESULTS ARE IMPROVED FURTHER IF MODELS ARE FIRST TRAINED ON NON-DISASTER-HIT IMAGES FROM THE SAME DATASET (*xBD-pre*) AND THEN EVALUATED AT DIFFERENT LOCATIONS DEPICTING DISASTER-HIT SCENARIOS.

Agent type	Success	Steps	Runtime	Success	Steps	Runtime
AiRLoc (5x5)	66.1 %	6.5	130 ms	72.8 %	6.1	122 ms
Priv local (5x5)	63.8 %	6.7	121 ms	67.3 %	6.4	115 ms
Priv random (5x5)	40.8 %	7.9	48 ms	40.8 %	7.9	48 ms
AiRLoc (7x7)	50.7 %	10.2	204 ms	55.7 %	9.9	198 ms
Priv local (7x7)	50.5 %	10.2	184 ms	53.6 %	10.0	180 ms
Priv random (7x7)	25.5 %	12.2	74 ms	25.5 %	12.2	74 ms

RL-vision setups [25]. As seen in Table II, AiRLoc and the best alternative learnable approach *Priv local* generalize excellently to an entirely new dataset.

Table III contains results on *xBD-disaster*; these results are particularly relevant from a perspective of SAR-operations in disaster-hit areas. Columns 1-3 show that AiRLoc generalizes quite well from having been trained on an entirely different dataset (*Masa*), which depicts non-disaster-hit urban areas, to disaster-hit areas at various other spatial locations. Results are however improved further (columns 4-6) if models are first trained on non-disaster-hit images from the same dataset (*xBD-pre*) and then evaluated at different locations that depict disaster-hit scenarios.

In summary, AiRLoc outperforms the baselines across all datasets and search area sizes, and localizes goals in fewer steps on average. See Fig. 2 and Fig. 5 - 6 (the latter two are on the last page) for visualizations of AiRLoc and *Priv local*.

Human performance evaluation. The results of the proof-of-concept human performance evaluation in Fig. 3 (left) indicate that our proposed task is in general difficult, since only slightly above half of all human controlled trajectories are successful. We also see that AiRLoc and *Priv local* achieve significantly higher success rates compared to human operators. Details about the human performance evaluation are found in the appendix.

B. Ablation Study: Motivating the Exploitation Prior

In Fig. 3 we evaluate the various AiRLoc variants described earlier, together with the best non-RL-based model *Priv local* and the human baseline. AiRLoc is better than its ablated variants on average in both settings (5×5 and 7×7), as well as for most start-to-goal distances (exception at distance 4 in the 7×7 setting). This motivates the design choice of fully utilizing the exploitation prior within the policy architecture – see also Table IV.

Recall that *Priv local* is trained solely in the setting where the start and goal are adjacent, so it can be interpreted as an 'exploitation only' model, where the action distribution is obtained by feeding the exploitation prior $\boldsymbol{u}_t$ through a softmax, cf. Fig. 1. Conversely, the *no prior* variant of AiRLoc is trained without any exploitation prior, so the policy must simultaneously learn to explore (search for the goal when it is further away) and exploit (move to the goal when it is adjacent), which may be ambiguous. As seen in

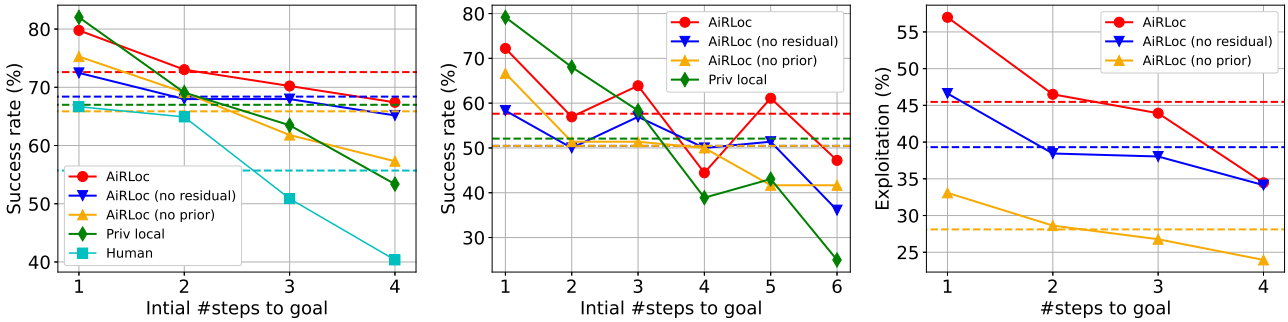

Fig. 3. Left and mid: Success rate versus start-to-goal distance on the validation set of *Masa* (averages are dashed). Search areas are of size $M \times N = 5 \times 5$ and $T = 10$ (left) or 7×7 and 14 (middle). Left: The methods are generally more successful when the start is closer to the goal. AiRLoc and *Priv local* achieve higher success rates than human operators. AiRLoc performs roughly on par with *Priv local* when the goal and start are adjacent (*Priv local* is trained only in this setting) and outperforms it at larger distances. AiRLoc is also more successful than its ablated variants in all settings. Middle: AiRLoc is best on average, despite having only been trained in the 5×5 setting. *Priv local* is better when the start and goal are close to each other, while AiRLoc is better when they are three or more steps apart. Right: How frequently AiRLoc selects the same action as the exploitation prior (argmax of $\boldsymbol{u}_t$) versus goal distance. The full AiRLoc agent has the largest variability in exploitation versus exploitation depending on distance to goal.

Fig. 3, the *no residual* variant, which allows $\boldsymbol{u}_t$ to guide the agent's decision making by feeding $\boldsymbol{u}_t$ to the temporal unit, is only marginally better. Our full AiRLoc agent, which clearly outperforms the other variants, takes this a step further by decoupling exploration and exploitation and only has to learn a residual between the two (since $\boldsymbol{u}_t$ is added within the softmax of the decision unit). Hence, during RL training AiRLoc essentially learns when to explore and when to exploit.

C. Random Seed Sensitivity Analysis

Table IV shows the results of a seed sensitivity analysis (regarding policy network initilization) for AiRLoc and its ablated variants on the validation set of *Massachusetts Buildings*. The AiRLoc variants are trained with five random network initializations each until convergence on the validation set, and the results for the median-performing models on the validation set are the ones reported within the rest of the paper. The seed sensitivity is low overall. Furthermore, our full AiRLoc agent outperforms *Priv local* even for the worst-performing seed.

V. Conclusions

In this work we have introduced the novel *aerial view goal localization* task and framework, which allows for controllable and reproducible development of methodologies that can eventually be useful for automated search-and-rescue operations, e.g. in regions that are heavily affected by climate-induced disasters. Naturally, as with most technologies, there are also possible applications that may be unethical. We strongly discourage extending our research in such directions, and instead call for extensions towards benign use-cases.

The difficulty for humans to perform well on our proposed task shows that it is a reasonable first step for model development and evaluation, even though the setup avoids some challenges of real use-cases. Relevant next steps toward making the proposed methodologies more practically useful include making the goal specification more flexible (e.g. allowing for a ground-level image description of the goal); requiring the agent to explicitly declare when it has reached its goal; and considering even larger search areas.

An RL-based approach, *AiRLoc*, was developed to tackle the proposed task, in addition to several other learnable and heuristic methods. Key components of the policy architecture include a mid-level vision module and an explicit decoupling between exploration and exploitation, both of which were shown to be crucial for AiRLoc's performance. Extensive experimental evaluations clearly showed the benefits of our AiRLoc agent over the learnable and heuristic baselines. In particular, our methodology can be used to localize goals in aerial images depicting disaster zones, despite being trained only on scenarios without disasters. Code and models have been made publicly available[3] so that others can further explore and extend our proposed task towards real use-cases, for example within disaster relief and management.

Appendix

In this appendix we provide further details about the human performance evaluation. To compare the performance of AiRLoc with a human operator in a similar setting, a game version of the task was developed. For fair comparisons, this game was designed to resemble how AiRLoc perceives the search area. Therefore, in addition to receiving the current and goal patches, the human operator is also aware of the borders of the search area, and knows the current position as well as the history of all previously visited positions within the confined area – see Fig. 4. In fact, the human operator can even see all the previously visited patches, while this information is not provided to AiRLoc. We decided to provide humans with this additional information as they have not been trained for the task at hand. Based on this input, the human operator can move to any of the eight adjacent patches. The movement is selected by clicking with a mouse cursor on one of the eight dark squares surrounding the current location in the *Player Area*, shown on the left in Fig. 4. The game uses search areas of size 5×5 and ends either when the movement budget $T = 10$ is exhausted or

[3] https://github.com/aleksispi/airloc

TABLE IV

SEED SENSITIVITY ANALYSIS OF THE VARIOUS AIRLOC VARIANTS ON THE VALIDATION SET OF *Massachusetts Buildings* (SEARCH AREA SIZE 5×5, MOVEMENT BUDGET $T = 10$). THE RESULTS ON THE FIRST LINES OF EACH BLOCK ARE THE MEDIAN-PERFORMING AIRLOC MODELS AND ARE THE ONES WE HAVE EVALUATED IN THE REST OF THE PAPER. NONE OF THE AIRLOC VARIANTS ARE SENSITIVE TO THE RANDOM SEED USED FOR POLICY NETWORK INITIALIZATION. THE WORST PERFORMING SEED OF THE *no residual* VARIANT OF AIRLOC PERFORMS BETTER THAN THE BEST PERFORMING SEED OF THE *no prior* VARIANT, AND IT IS ALSO SOMEWHAT BETTER THAN THE ALTERNATIVE LEARNABLE APPROACH *Priv local*. SIMILARLY, THE WORST PERFORMING SEED OF OUR FULL AIRLOC OUTPERFORMS THE BEST PERFORMING SEED OF BOTH THE ABLATED VARIANTS AND *Priv local*, WHICH AGAIN MOTIVATES OUR DESIGN CHOICES.

Agent type	Success	Step ratio	Steps	Residual distance
AiRLoc	72.6 %	1.49	6.0	2.4
AiRLoc (other seed #1)	72.2 %	1.45	6.1	2.4
AiRLoc (other seed #2)	72.2 %	1.51	6.2	2.5
AiRLoc (other seed #3)	74.3 %	1.56	6.2	2.4
AiRLoc (other seed #4)	75.9 %	1.53	6.1	2.5
AiRLoc (average)	73.4 %	1.51	6.1	2.5
AiRLoc (no residual)	68.5 %	1.49	6.3	2.2
AiRLoc (no residual, other seed #1)	68.6 %	1.52	6.3	2.2
AiRLoc (no residual, other seed #2)	69.5 %	1.52	6.3	2.2
AiRLoc (no residual, other seed #3)	68.2 %	1.60	6.4	2.2
AiRLoc (no residual, other seed #4)	67.2 %	1.57	6.4	2.2
AiRLoc (no residual, average)	68.4 %	1.54	6.3	2.2
AiRLoc (no prior)	65.9 %	1.56	6.5	2.4
AiRLoc (no prior, other seed #1)	64.8 %	1.56	6.7	2.4
AiRLoc (no prior, other seed #2)	66.6 %	1.56	6.5	2.5
AiRLoc (no prior, other seed #3)	66.6 %	1.50	6.4	2.3
AiRLoc (no prior, other seed #4)	64.9 %	1.50	6.6	2.4
AiRLoc (no prior, average)	65.8 %	1.54	6.5	2.4
Priv local	67.0 %	1.54	6.3	2.3

when the player moves into the goal location (just as for AiRLoc and the other baselines). Moreover, different to the other approaches, the human participants have a limited time to complete each game (60 seconds). Such a time limit was used for the convenience of the participants – we wanted to avoid that the participants felt like they had to spend several minutes per action to squeeze out the maximum possible performance. The 60 second time limit was assessed to be more than sufficient for completing each game, and the participants agreed with this.

The age span of the 19 people who participated is between 14 and 42 years, with an average of 26.4 years and a median of 25 years. There were 13 men and 6 women (68% and 32%, respectively). For each human operator, 12 unique search areas from the validation set of *Massachusetts Buildings* were used, as well as a few sample search areas for the player to get acquainted with the controls of the game – the participants were able to practice as long as they desired, and no statistics were tracked during this warm up phase. The exact games provided span a subset of the games that AiRLoc and the other baselines are evaluated on, to ensure that the comparison is as fair as possible. However, each human is not tested on the entire dataset since it is impractically large, and hence there is a higher uncertainty in the human performance evaluation. The difficulty settings were split equally over these twelve games, with three games per difficulty (here difficulty is the distance between the start and goal patches, ranging from 1 to 4 steps away).

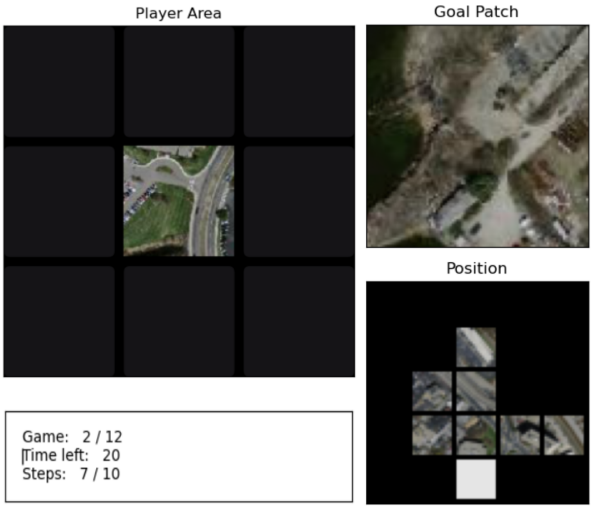

Fig. 4. An example of the human performance evaluation setup. Each participant was given a set of 12 different such games (a game is a search area and an associated start and goal location), and there was no overlap in the games played by different participants. Each search area was of size 5×5 and the movement budget was $T = 10$.

Even though the human setup is very similar to that of AiRLoc, there are some concepts that do not translate well to a human controlled setup. First, the positional encoding of AiRLoc is difficult to translate to human visual processing, and instead a map of the positions was implemented (thus the participants receive explicit information from past locations,

different from AiRLoc). Second, the human participants have not trained on the task like AiRLoc, and their visual systems are likely not tailored towards handling the quite low resolution patches. On the other hand, humans have implicitly conducted a lifetime worth of generic visual pretraining, which AiRLoc has not. These discrepancies, in conjunction with the limited number of human controlled trajectories, somewhat limit the reliability of the human baseline. Nonetheless, it is still a useful indication of the human performance on our proposed task.

REFERENCES

[1] Peter Anderson, Angel Chang, Devendra Singh Chaplot, Alexey Dosovitskiy, Saurabh Gupta, Vladlen Koltun, Jana Kosecka, Jitendra Malik, Roozbeh Mottaghi, Manolis Savva, et al. On evaluation of embodied navigation agents. *arXiv preprint arXiv:1807.06757*, 2018.

[2] Kumar Ayush, Burak Uzkent, Kumar Tanmay, Marshall Burke, David Lobell, and Stefano Ermon. Efficient poverty mapping using deep reinforcement learning. *arXiv preprint arXiv:2006.04224*, 2020.

[3] Luca Bartolomei, Lucas Teixeira, and Margarita Chli. Perception-aware path planning for uavs using semantic segmentation. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2020.

[4] Adrian Boguszewski, Dominik Batorski, Natalia Ziemba-Jankowska, Tomasz Dziedzic, and Anna Zambrzycka. Landcover.ai: Dataset for automatic mapping of buildings, woodlands, water and roads from aerial imagery, 2020.

[5] Juan C Caicedo and Svetlana Lazebnik. Active object localization with deep reinforcement learning. In *Proceedings of the IEEE international conference on computer vision*, pages 2488–2496, 2015.

[6] Tung Dang, Christos Papachristos, and Kostas Alexis. Autonomous exploration and simultaneous object search using aerial robots. In *2018 IEEE Aerospace Conference*, pages 1–7. IEEE, 2018.

[7] Carl Doersch, Abhinav Gupta, and Alexei A. Efros. Unsupervised visual representation learning by context prediction, 2015.

[8] Mingfei Gao, Ruichi Yu, Ang Li, Vlad I Morariu, and Larry S Davis. Dynamic zoom-in network for fast object detection in large images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6926–6935, 2018.

[9] Ritwik Gupta, Bryce Goodman, Nirav Patel, Ricky Hosfelt, Sandra Sajeev, Eric Heim, Jigar Doshi, Keane Lucas, Howie Choset, and Matthew Gaston. Creating xbd: A dataset for assessing building damage from satellite imagery. In *CVPR workshops*, 2019.

[10] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9:1735–80, 12 1997.

[11] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015.

[12] Danil Kuzin, Olga Isupova, Brooke D Simmons, and Steven Reece. Disaster mapping from satellites: damage detection with crowdsourced point labels. *arXiv preprint arXiv:2111.03693*, 2021.

[13] Ajith Anil Meera, Marija Popović, Alexander Millane, and Roland Siegwart. Obstacle-aware adaptive informative path planning for uav-based target search. In *ICRA*, 2019.

[14] Lina Mezghani, Sainbayar Sukhbaatar, Thibaut Lavril, Oleksandr Maksymets, Dhruv Batra, Piotr Bojanowski, and Karteek Alahari. Memory-Augmented Reinforcement Learning for Image-Goal Navigation. working paper or preprint, March 2022.

[15] Volodymyr Mnih. *Machine Learning for Aerial Image Labeling*. PhD thesis, University of Toronto, 2013.

[16] Volodymyr Mnih, Nicolas Heess, Alex Graves, et al. Recurrent models of visual attention. *NeurIPS*, 2014.

[17] Simone Parisi, Aravind Rajeswaran, Senthil Purushwalkam, and Abhinav Gupta. The unsurprising effectiveness of pre-trained vision models for control. *arXiv preprint arXiv:2203.03580*, 2022.

[18] Aleksis Pirinen and Cristian Sminchisescu. Deep reinforcement learning of region proposal networks for object detection. In *CVPR*, 2018.

[19] Marija Popović, Teresa Vidal-Calleja, Gregory Hitz, Jen Jen Chung, Inkyu Sa, Roland Siegwart, and Juan Nieto. An informative path planning framework for uav-based terrain monitoring. *Autonomous Robots*, 44(6):889–911, 2020.

[20] Shraman Pramanick, Ewa M Nowara, Joshua Gleason, Carlos D Castillo, and Rama Chellappa. Where in the world is this image? transformer-based geo-localization in the wild. *arXiv preprint arXiv:2204.13861*, 2022.

[21] Samrudhdhi B Rangrej and James J Clark. A probabilistic hard attention model for sequentially observed scenes. *arXiv preprint arXiv:2111.07534*, 2021.

[22] Samrudhdhi B Rangrej, Chetan L Srinidhi, and James J Clark. Consistency driven sequential transformers attention model for partially observable scenes. *arXiv preprint arXiv:2204.00656*, 2022.

[23] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation, 2015.

[24] Seyed Abbas Sadat, Jens Wawerla, and Richard Vaughan. Fractal trajectories for online non-uniform aerial coverage. In *ICRA*, 2015.

[25] Alexander Sax, Bradley Emi, Amir R Zamir, Leonidas Guibas, Silvio Savarese, and Jitendra Malik. Mid-level visual representations improve generalization and sample efficiency for learning visuomotor policies. *arXiv preprint arXiv:1812.11971*, 2018.

[26] Michael Schmitt, Pedram Ghamisi, Naoto Yokoya, and Ronny Hänsch. Eod: The ieee grss earth observation database. In *IGARSS 2022-2022 IEEE International Geoscience and Remote Sensing Symposium*, pages 5365–5368. IEEE, 2022.

[27] Felix Stache, Jonas Westheider, Federico Magistri, Cyrill Stachniss, and Marija Popović. Adaptive path planning for uavs for multi-resolution semantic segmentation. *arXiv preprint arXiv:2203.01642*, 2022.

[28] Humans In the Loop. Semantic segmentation of aerial imagery.

[29] Burak Uzkent and Stefano Ermon. Learning when and where to zoom with deep reinforcement learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12345–12354, 2020.

[30] Andrea Vallone, Frederik Warburg, Hans Hansen, Søren Hauberg, and Javier Civera. Danish airs and grounds: A dataset for aerial-to-street-level place recognition and localization. *CoRR*, abs/2202.01821, 2022.

[31] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2017.

[32] Che Wang, Xufang Luo, Keith Ross, and Dongsheng Li. Vrl3: A data-driven framework for visual deep reinforcement learning. *arXiv preprint arXiv:2202.10324*, 2022.

[33] Tingyu Wang, Zhedong Zheng, Yaoqi Sun, Tat-Seng Chua, Yi Yang, and Chenggang Yan. Multiple-environment self-adaptive network for aerial-view geo-localization. *arXiv preprint arXiv:2204.08381*, 2022.

[34] R. J. Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*, 8:229–256, 1992.

[35] Daniel Wilson, Xiaohan Zhang, Waqas Sultani, and Safwan Wshah. Visual and object geo-localization: A comprehensive survey. *arXiv preprint arXiv:2112.15202*, 2021.

[36] Junshi Xia, Naoto Yokoya, Bruno Adriano, and Clifford Broni-Bediako. Openearthmap: A benchmark dataset for global high-resolution land cover mapping. *arXiv preprint arXiv:2210.10732*, 2022.

[37] Tete Xiao, Ilija Radosavovic, Trevor Darrell, and Jitendra Malik. Masked visual pre-training for motor control. *arXiv preprint arXiv:2203.06173*, 2022.

[38] Zhitong Xiong, Fahong Zhang, Yi Wang, Yilei Shi, and Xiao Xiang Zhu. Earthnets: Empowering ai in earth observation. *arXiv preprint arXiv:2210.04936*, 2022.

[39] Karmesh Yadav, Ram Ramrakhya, Arjun Majumdar, Vincent-Pierre Berges, Sachit Kuhar, Dhruv Batra, Alexei Baevski, and Oleksandr Maksymets. Offline visual representation learning for embodied navigation. *arXiv preprint arXiv:2204.13226*, 2022.

[40] Zelong Zeng, Zheng Wang, Fan Yang, and Shin'ichi Satoh. Geo-localization via ground-to-satellite cross-view image retrieval. *IEEE Transactions on Multimedia*, pages 1–1, 2022.

[41] Leyang Zhao, Li Yan, Xiao Hu, Jinbiao Yuan, and Zhenbao Liu. Efficient and high path quality autonomous exploration and trajectory planning of uav in an unknown environment. *ISPRS International Journal of Geo-Information*, 10(10):631, 2021.

[42] Runzhe Zhu. Sues-200: A multi-height multi-scene cross-view image benchmark across drone and satellite, 2022.

[43] Yuke Zhu, Roozbeh Mottaghi, Eric Kolve, Joseph J Lim, Abhinav Gupta, Li Fei-Fei, and Ali Farhadi. Target-driven visual navigation in indoor scenes using deep reinforcement learning. In *ICRA*, 2017.

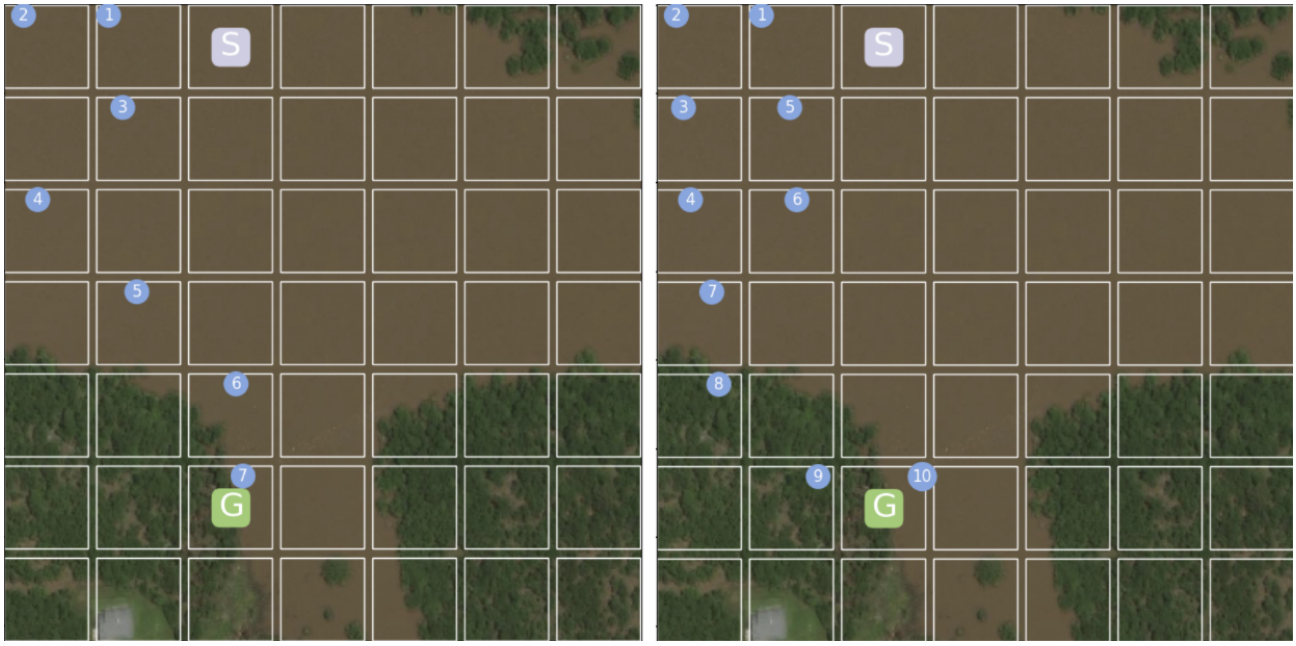

Fig. 5. Successful examples of AiRLoc (left) and *Priv local* (right) on a flooding scenario in *xBD-disaster* (7×7 setup, movement budget $T = 14$). The start and goal locations are denoted 'S' and 'G', respectively. The numbered circles show which locations are visited and in what order. Recall that the full underlying search area is never observed in its entirety, i.e. the agents must operate based on partial glimpses alone. Also note that AiRLoc was only trained on search areas of size 5×5 and movement budget $T = 10$. AiRLoc takes the same first two steps as *Priv local*, then deviates and reaches the goal in fewer steps than *Priv local*.

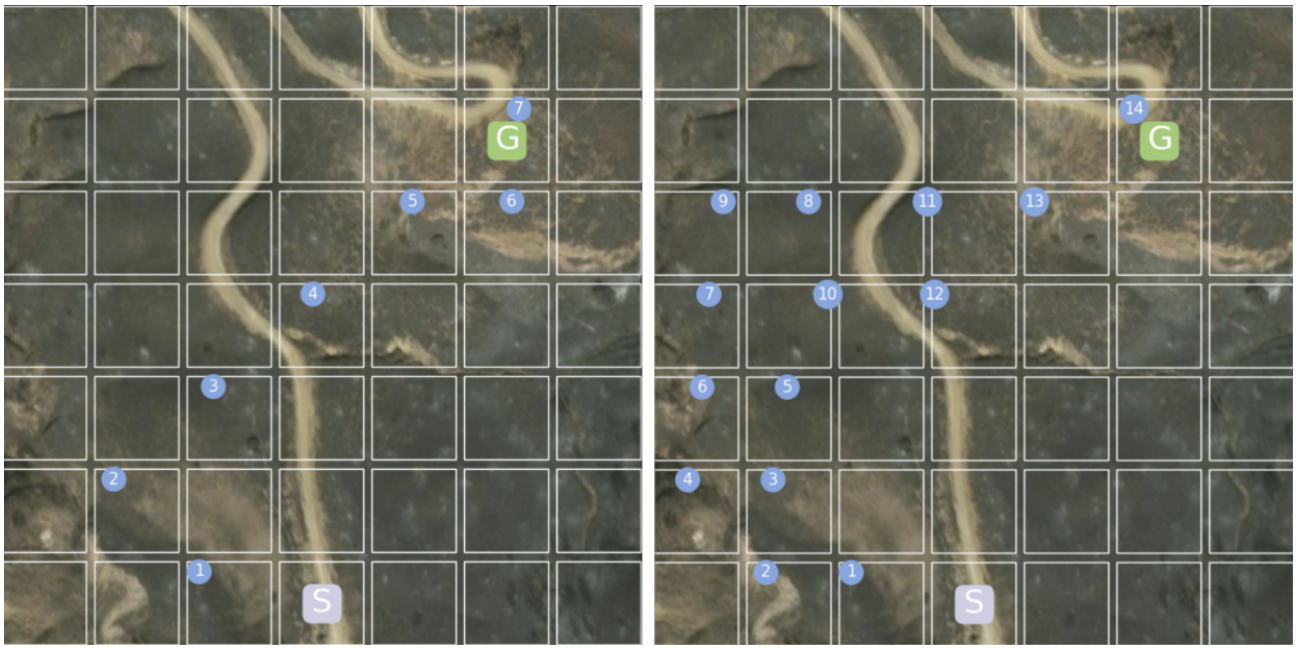

Fig. 6. Successful examples of AiRLoc (left) and *Priv local* (right) on a post-wildfire scenario in *xBD-disaster* (7×7 setup, movement budget $T = 14$). AiRLoc takes the same first step as *Priv local*, then deviates, and reaches the goal twice as fast. *Priv local* precisely manages to reach the goal within its movement budget.

Urdarbrunnen: Towards an AI-enabled mission system for Combat Search and Rescue operations

Ella Olsson[1,*] Mikael Nilsson[2,*] Kristoffer Bergman[3,*] Daniel de Leng[2,4,*]
Stefan Carlén[2] Emil Karlsson[2] Bo Granbom[2,†]

***Abstract*— The Urdarbrunnen project is a Saab-led exploratory initiative that aims to develop an operator-assisted AI-enabled mission system for basic autonomous functions. In its first iteration, presented in this project paper, the system is designed to be capable of performing the search task of a combat search and rescue mission in a complex and dynamic environment, while providing basic human machine interaction support for remote operators. The system enables a team of agents to cooperatively plan and execute a search mission while also interfacing with the WARA-PS core system that allows human operators and other agents to monitor activities and interact with each other. The aim of the project is to develop the system iteratively, with each iteration incorporating feedback from simulations and real-world experiments. In future work, the capability of the system will be extended to incorporate additional tasks for other scenarios, making it a promising starting point for the integration of autonomous capabilities in a future air force.**

I. Introduction

Having rapidly progressed from expensive and custom-made hardware solutions to commercially-available off-the-shelf products, unmanned aerial vehicles (UAVs), colloquially referred to as 'drones', are becoming an increasingly common sight in today's airspace. These vehicles are often part of unmanned aircraft systems (UAS) that can be used for a wide variety of tasks, both civil and military, ranging from camera shoots for movies and consumer product deliveries to improvised or specialized weapon platforms as observed in the Russian Federation's ongoing invasion of Ukraine. The latter has shown that these type new technologies are crucial in today's combat environment which motivates further long-term investments in innovative research [1].

The prevalence of UAVs in civilian applications is also resulting in the evolution of airspace management, with the European Union's U-Space airspace initiative [2] expected to come into effect soon—thereby opening up novel opportunities for European business sectors—and autonomous airport solutions to accommodate them. On the other hand, the ease with which private actors can acquire and operate UAVs has resulted in new security challenges that also affect protected areas of national interest such as airports and power plants.

[1]Saab AB, Nettovägen 6, SE-175 41 Järfälla, Sweden.
[2]Saab AB, Bröderna Ugglas Gata, SE-581 88 Linköping, Sweden.
[3]RISE Research Institutes of Sweden AB, Fridtunagatan 41, SE-582 16 Linköping, Sweden.
[4]Department of Computer and Information Science, Linköping University, SE-581 83 Linköping, Sweden.
*These authors contributed equally.
† Corresponding author: bo.granbom@saabgroup.com

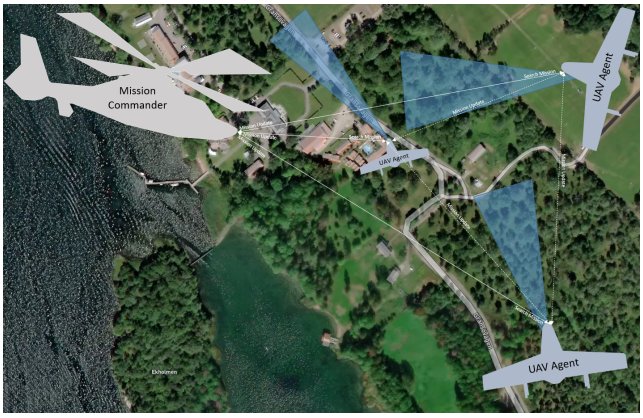

Fig. 1. An illustration of actors in a Combat Search and Rescue scenario, including a mission commander and a number of autonomous UAV agents performing a collaborative search. The background environment in the figure is courtesy of WARA-PS.

Saab is a Swedish security and defense company that strives to keep society and people safe [3]. In furtherance of this goal, Saab is the producer of defense and security products and services, including Saab's Gripen fighter jet [4].

In this document we use the term Tactical Autonomy which we define as a technology that relate to functions that jointly and independently aim to fulfil a mission goal through the selection of different courses of action based on intrinsic knowledge and understanding of the situation and itself, as well as the predicted outcomes and associated constraints, such as risk acceptance, available resources, etc. Collaboration and teaming with human operators, as well as with other autonomous functions (multi-agent collaboration) is an essential part of the technology area.

There is an emergence of tactical autonomy solutions within the air domain, and in particular for UAS. It is envisioned that these solutions have the potential to drastically change the way security and defense are ensured. These systems have the potential of acting as a force multiplier in the areas of security and defense, especially in mixed teams consisting of autonomous agents and skilled humans. This is not a new idea either; UAS have already been demonstrated to be useful in civilian search and rescue (SAR) scenarios, with the *Hybrid Deliberative/Reactive* (HDRC3) framework [5] and related *WASP Research Arena for Public Safety* (WARA-PS) Core System [6] being among the pioneering research in that area.

This paper presents *Urdarbrunnen*[5], which is an exploratory effort led by Saab towards developing a framework that can coordinate missions, i.e. a mission system, involving manned and unmanned systems operating in the air domain, with the goal of better understanding and learning about what kind of frameworks may be needed in the future. Concretely, this paper focuses on the first iteration of a planning, coordination and execution framework for mixed manned/unmanned missions. The overarching goal of the Urdarbrunnen project is to develop an initial architectural design that can be implemented and integrated in commercial off-the-shelf remotely piloted aircraft systems to provide full autonomy comparable to HDRC3, but in an operational domain in which it is able to complement and augment the capabilities of a modern air force. Following the example of SAR missions in the WASP Research Arena for Public Safety [6], we will focus on combat search and rescue (CSAR) missions in which a mission commander and autonomous UAV agents take on a (non-combat) supporting role; see Figure 1. CSAR missions differentiate themselves from SAR missions in a number of ways, including in the potential presence of an adversarial and disruptive force element, which make them a better fit for Urdarbrunnen.

The remainder of this paper is organized as follows. In Section II, we consider the place of autonomous UAS solutions in the military domain. Section III then presents the CSAR scenario which forms the basis of the Urdarbrunnen architecture design. The mission planning system underlying Urdarbrunnen is discussed in more detail in Section IV. An instantiation of the Urdarbrunnen architecture on a UAV Agent is presented in Section V, which is followed by a CSAR mission example in Section VI. We conclude the paper and look towards future work in Section VII.

II. Background

Contemporary autonomous systems depend on cognitive capabilities to monitor their surroundings, estimate the current state of the world, and predict what might happen next. They can make use of artificial intelligence (AI) algorithms and models in order to reason about, learn from, and interact with the world. Autonomous systems have been the subject of research and development activities globally for decades, at varying levels of complexity, and are consistently mentioned in strategy documents. For example, in 2005 the United States Department of Defense issued the "Unmanned Aircraft Systems Roadmap 2005–2030" [7], in which UAS are expected to possess various levels of autonomy. In addition, the Swedish commander-in-chief of the armed forces has expressed [8] an urgent need for e.g. autonomous capability development in the light of a coming NATO membership and current security policies.

Autonomous unmanned systems have a number of advantages over conventional manned systems or even remotely-piloted unmanned systems. They are 1) often smaller, allowing for operations in areas that are unsuitable for manned platforms, 2) usually cheaper to develop, operate, and maintain, and 3) extending the operational domain to areas that are unreachable or unsuitable for manned platforms, hence can be used in situations that would otherwise be deemed too risky to a pilot or operator [9], [10]. Common tasks for these systems include supporting personnel on the ground or in the air through the delegation of high-level tasks, where AI methods are commonly used to break down and execute these tasks. Crucially, the collaboration between autonomous unmanned systems adds an additional layer of capabilities that utilize teams of systems of systems. This collaboration can be exemplified by multiple autonomous unmanned systems negotiating plans towards meeting an operator-set goal under a set of provided constraints. One example of a civilian system that is capable of collaborative planning is the HDRC3 system, which is used within both the WARA-PS research arena [6] and within the Autonomous Search System (AuSSys) [11] research project. Another example of a control architecture for controlling multiple UAVs for SAR in alpine scenarios is given in [12], where one human operator is able to coordinate the actions of multiple UAVs. A more detailed overview of recent work within AI-based mission planning for unmanned vehicles is found in the survey paper [13]. In the context of this paper, we instead focus on the military domain.

III. Scenario

Many research results employ a Search and Rescue SAR style scenarios to demonstrate novel capabilities. In a similar vein, we adopt the CSAR [14] task and we let this task inspire our scenario of interest.

A. Mission

As described in the United States' CSAR Air Force Doctrine [14], a successful CSAR operation "enhances the Joint Force Commander's (JFC) combat capability by returning personnel to areas under friendly control and denying adversaries the opportunity to exploit the intelligence and propaganda value of captured personnel.". Whereas the primary objective of a CSAR mission is to recover isolated personnel such as downed aircrew, we will in the first phase of this project mainly focus on the initial phases where determining the location(s) of isolated personnel in an adversarial environment is of key importance. Therefore our main operational scenario focuses on the search part of the CSAR task, and we aim to employ search-capable fixed-wing UAVs for the search sub-task.

The scenario preparation consists of decomposing the CSAR task into a set of sub-tasks, namely a *Search*, a *Rescue* and a *Combat* sub-task—but as stated above, the focus in this development phase is on the Search sub-task. We situate the area of operations of this sub-task in the vicinity of Gränsö Castle, in Västervik motivated by the excellent research and demonstration opportunities available at this location, as well as the opportunity of being a part of the multi-domain community within the WARA-PS research program. WARA-PS [6] also offers the possibility to conduct mixed

[5]Based on Norse mythology: The Norns living near Urdarbrunnen were thought to determine a person's fate.

academic/industrial research in multiple research areas, including but not limited to command and control of UAS in Search and Rescue missions, as well as offering excellent demonstration facilities in the Västervik area where research results can be presented and demonstrated. Therefore we align the search scenario—including its general environment where the scenario is situated in terms of i.e. topography, features, weather—with this location. This environment can be defined as a mix of open grassy terrain, leaf vegetation, shoreline and open water. We partly include open water for our task as the search operation may transverse from land to open water during the search. In the first phase of this project, we assume that the electromagnetic environment is non-obstructed, allowing us to exercise our communication links at full capacity. Furthermore, in this phase, we also assume that there are no hostile signal intelligence or communications intelligence present restricting the communication in relation of transmission and confidentiality aspects. Target intelligence includes one or more persons in distress, located anywhere in the vicinity of Gränsö castle, and we assume that this scenario instance (albeit unbeknownst to the agents) does not include any hostile agent threats, as the initial phase of this project mainly will focus on determining the isolated person(s) location. We aim to include threats in later phases of the project where we also will focus on the rescue and the combat support effort.

With regards to other intelligence objects we include our home base as a position for take-off and landing of our resources. In terms of cooperative forces, we include the ability to cooperate with external systems in the land, the maritime and the air domain. The purpose of this is to increase the operational effectiveness of our search operation. It might also exist neutral entities in the scenario that we must consider for safety reasons. Neutral entities may be civil boats or other vehicles, people, wild life etc. Our resources consists of a team of fixed-wing UAVs, each equipped with a pedestal mounted gimbal camera.

B. Measures

We develop our version of the CSAR mission influenced by the CSAR task as defined in the United States Air Force Task List (AFTL) [15]. The task is listed under the capability PROVIDE PRECISION ENGAGEMENT within the framework for expressing the Air Force tasks, where PROVIDE PRECISION ENGAGEMENT is defined as "to command, control, and employ forces to cause discriminate strategic, operational or tactical effects." [15, p. 87]. The CSAR task itself is described to includes capabilities "to organize, train, equip, provide, and plan for the conduct of prompt and sustained air operations to recover isolated personnel during wartime and contingency operations." [15, p. 90]. Within the CSAR task we focus on the on two CSAR functions in particular:

- AFT 2.3.1 PERFORM CSAR FUNCTIONS: "To conduct operations to recover isolated personnel during wartime or contingency as necessary." [15, p. 90].
- AFT 2.3.4 PLAN CSAR FUNCTIONS: "To consider all the particulars associated with the optimum utilization of CSAR resources and to produce the necessary products to ensure effectiveness of CSAR functions is maximized." [15, p. 91].

TABLE I
BREAKDOWN OF PERFORM CSAR FUNCTIONS.

AFT 2.3.1 PERFORM CSAR FUNCTIONS "To conduct operations to recover isolated personnel during wartime or contingency as necessary."		
M1	Time	to recover distressed isolated personnel during wartime or contingency as necessary.
M2	Number	of personnel recovered during wartime or contingency operations.
M3	Percent	of successful CSAR operations.
M4	Cost	to perform CSAR functions.

TABLE II
BREAKDOWN OF PLAN CSAR FUNCTIONS.

AFT 2.3.4 PLAN CSAR FUNCTIONS "To consider all the particulars associated with the optimum utilization of CSAR resources and to produce the necessary products to ensure effectiveness of CSAR functions is maximized."		
M1	Percent	of resources used to conduct CSAR functions properly planned.
M2	Percent	of shortcomings in plans used to conduct CSAR functions.
M4	Time	to complete required planning to conduct CSAR functions.
M5	Cost	to plan CSAR functions.

Based on these functions we develop the CSAR agent architecture as defined in and detailed in sections IV and V. We also adopt the corresponding measurements on a functional level as defined in the AFTL [15, p. 90-91] for these functions in order to perform an adequate evaluation of the mission. Inspired by the AFTL, we have selected the CSAR task as a *Mission Essential Task*. This has helped us to determine *what* to do, i.e. plan and execute the Search-part of a CSAR mission. We have also determined the conditions for this task by means of a scenario definition as detailed in Section III. The final step in developing our mission requirements involves selecting performance measures for the CSAR task as described in the AFTL [15, p. 64]. In this development phase, we omit the establishment of standards as also described in the AFTL [15, p. 64] as we at the moment of writing this paper, do not have the minimum acceptable proficiency required performance for the task at hand. The specific measures are selected from the AFTL [15, p. 90-91] and are detailed in Tables I and II.

IV. URDARBRUNNEN PLANNING SYSTEM

In order for a system of autonomous agents to achieve complex goals, coordinated planning and execution of plans are essential capabilities. Autonomous agents must able to handle unexpected events during plan execution. Together these aspects require a tight coupling between planning and execution. It also requires any participating autonomous agent to be able to perform at least rudimentary local planning as far as it itself is concerned.

Planning in autonomous agents is facilitated by automated planning. Automated planning is a rich field within AI that

over the years has provided many different approaches to planning in many different domains. Research in automated planning has led to the development of a common planning language named *Planning Domain Definition Language* (PDDL) [16], [17], and many of the various planners available support planning in domains that make use of a subset of this language. Planners that can derive plans in any given domain are called *domain independent task planners*. They are often contrasted with *domain specific planners* that requires specific domain knowledge in order to plan in a given domain efficiently. The Urdarbrunnen planning system can leverage both types of planners depending on mission parameters.

Whenever agents interact it is important that their ontologies are aligned so that a concept like "flying" means the same to the agent performing the task and the agent planning it.

A. Abstraction level and planning approach

In order to provide a versatile architecture, capable of handling tasks of different complexity, the architecture should be capable of planning at different levels of abstractions. In a mission context, this naturally translates into being able to plan both centralized/globally and decentralized/locally. With centralized we mean that one planner derives the whole plan and with decentralized we mean that several planners provide smaller parts of a larger plan. Lower abstraction levels are suited for centralized planning and vice verse.

As an example of centralized low abstraction planning, a CSAR mission planner may plan almost every detail of each participant's actions, e.g. detailed commands for TAKEOFF, FLY-TO and LOOK-AT actions. But the mission could also be planned in a decentralized way at a higher level of abstraction, letting the top-level planner stop at the level of SEARCH-AREA commands, leaving the agents to perform the decentralized planning of how to SEARCH-AREA by themselves.

The architecture allows for different levels of abstraction depending on mission requirements and command preferences.

B. Initiatives – top-down vs bottom-up approach

When using a lower level of abstraction, the centralized planner makes all important decisions. This is a purely top-down approach where agents are left with less room to take initiative and have less responsibility. In a bottom-up approach, in contrast, a planner may break down missions into tasks and further into sub-tasks that are published and distributed to participating agents. Agents can then by their own initiative reserve and perform tasks, being fully responsible for carrying them out. In the bottom-up approach, agents are also responsible for synchronizing tasks among themselves. In order to facilitate publication, distribution and synchronization among agents we add a *Blackboard* and a *Constraint Store* to the architecture.

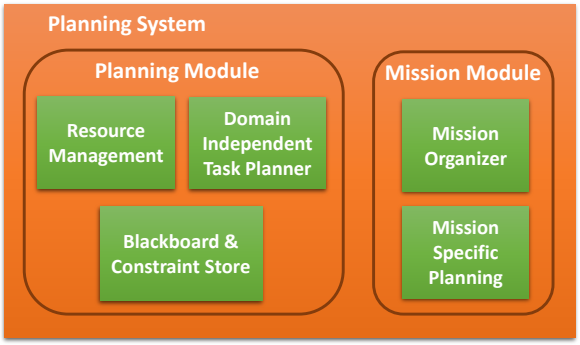

Fig. 2. Architecture of the Urdarbrunnen Planning System, consisting of a planning and a mission module.

C. Execution and re-planning

When a plan that solves the mission goal has been found, the next step is to execute it. In order to successfully execute a plan in the presence of unforeseen events, execution monitoring is needed. At the lowest level, an agent performing a task may need some kind of fallback, perhaps the execution follows a behavior tree model [18] that details what can go wrong and how to recover. If the agent cannot perform its task, plan execution enters a failure mode.

In a top-down architecture there is a global execution mechanism that when informed of the failure takes measures to repair the plan or come up with a new working plan.

A bottom-up architecture need to include a plan repair process that may involve first putting the failed task back on the blackboard and perhaps preventing the failing agent from reserving that task again after repeated failures.

D. The planning architecture

In order to meet the requirements discussed in previous sections, we define the planning system. An illustration of this system is found in Figure 2.

The planning system contains a mission independent *Planning Module* that contains the core planning capabilities that are needed to perform any mission. This is complemented by the mission-specific *Mission Module* that handles the mission-specific details for each type of mission that the system can perform.

The mission-independent planning modules are 1) *Resource Management*, keeping track of the systems resources and agent capabilities, 2) *Domain-Independent Task Planner*, required for planning and 3) *Blackboard & Constraint Store*, required to synchronize agents during missions.

The mission-dependent modules are the 1) *Mission Organizer*, responsible for putting together all aspects of the mission and executing it with the help of the 2) *Mission-Specific Planning*, containing all planning aspects that are outside of domain-independent task planning.

A concrete example of how to implement the planning system in a system capable of performing CSAR missions is given in the following sections.

V. URDARBRUNNEN UAV AGENTS

This section presents the system architecture for the Urdarbrunnen UAV agents in terms of hardware, middleware and software. An overview of an agent is shown in Figure 3. Note that the figure shows all software modules. It is possible for agents to be part of the system without having all modules.

A. Hardware

Autopilot: Pixhawk is an open-source hardware platform designed for the development of autonomous unmanned vehicles, such as drones, rovers, and other robotic platforms. It was first introduced in 2011 by the company 3D Robotics and has since become a popular choice among hobbyists, researchers, and commercial drone manufacturers. Pixhawk is compatible with various sensors, such as inertial measurement unit (IMU), Global Navigation Satellite Systems (GNSS), barometer and magnetometer, to provide a stable estimate of the physical state of the vehicle. The Pixhawk is also equipped with a micro controller that runs the firmware, which is responsible for controlling the motors, regulating the power supply, and communicating with other devices, such as a Ground Control Station (GCS).

Companion Computer: The UAV is also equipped with a companion computer. The companion computer is responsible for running the Robot Operating System 2 (ROS2) [19] software modules described in Section V-C, and is able to communicate with other UAV agents through the ROS2 network. The communication between the autopilot (Pixhawk) and the companion computer is done over Ethernet in order to minimize latency and maximize bandwidth.

RC Transmitter: The radio-control (RC) transmitter is used for remote control of the UAV by the safety pilot, whose main responsibility is to monitor flight and taking control of the vehicle if necessary to avoid any safety risks. Hence, the system must allow that a safety pilot is able to intervene.

B. Middleware

The UAV agents use ROS2 as a middleware in order to communicate and share information. ROS2 is a distributed and modular software framework designed for building robotic systems. The new version is designed to address some of the limitations of the original ROS framework, including limitations related to scalability, real-time performance, and support for various hardware platforms. ROS2 also incorporates new features and improvements, including support for multiple operating systems and programming languages, a more modular architecture, and better support for real-time and safety-critical applications. One of the main components of ROS2 is the communication layer based on the Data Distribution Service (DDS) standard [20].

C. Software

This section describes the Pixhawk autopilot software, as well as the different ROS2 modules running on the companion computer.

PX4: PX4 [21] is an open-source autopilot software developed specifically for the Pixhawk autopilots. It is a modular and highly configurable software stack that includes vehicle control, navigation, and mission planning functions. PX4 provides a flexible platform for the development and deployment of autonomous UAVs. It comes with support for various types of aircraft, including fixed-wing planes, multirotors, and Vertical TakeOff and Landing (VTOL) vehicles.

UAV Module: The UAV module is used to control the UAV and distribute information to other modules and agents. The bridging of messages between ROS2 and the PX4 software is done by connecting the microdds client in PX4 and a Micro XRCE-DDS Agent [22] in ROS2. The module contains the offboard flight controller, which bridges the flight control interface from the PX4 software into ROS2. The flight control component is thus responsible for exposing UAV-related ROS2 base actions such as TAKEOFF, LAND and FLY-TO. In swarm applications where tight coordinated control is required, such as formation flight, one could implement a flight control component in the UAV module that connects to and controls several UAVs simultaneously. The UAV module also contain components related to controlling the payload of the UAV, such as the camera. The camera control component exposes ROS2 APIs that can be used to perform actions such as TAKE-PHOTO, CONTROL-GIMBAL, RECORD-VIDEO and STREAM-VIDEO. Finally, the module contains a UAV information component, which main responsibility is to continuously distribute information about the state of the UAV (such as physical state, battery level, and flight-related information). The component is also aware of the different capabilities and attributes that are related to the UAV.

Information Module: The information module is used to collect and distribute all information that is relevant for the UAV agent. It provides a list of all capabilities and attributes that is associated with the UAV agent.

Planning and Execution Modules: Currently, the Urdarbrunnen UAV Agent uses the ROS2 based planning system PlanSys2 [23] to perform domain independent task planning. PlanSys2 enables easy handling of PDDL domains and problems by for instance facilitating incremental updates that reflect changes in the world. PlanSys2 is also capable of executing and monitoring the execution of derived plans. It relies however on external automated planners to do the actual planning. PlanSys2 is tested with the external planners POPF [24] and TFD [25], but any PDDL planner with the matching output format can be used, assuming it has a ROS2 integration. In the UAV Agent, the PlanSys2 PDDL executor is located in the Execution Module since this module can be used stand alone without the rest of the PlanSys2 system in a minimalist agent that relies on external planning. The rest of the PlanSys2 system is located in the PDDL Planner in the Planning Module.

The Resource Management module contains all available resources in the form of agent IDs and for each agent it also contains a list of capabilities belonging to it.

The Blackboard & Constraint Store is needed mainly

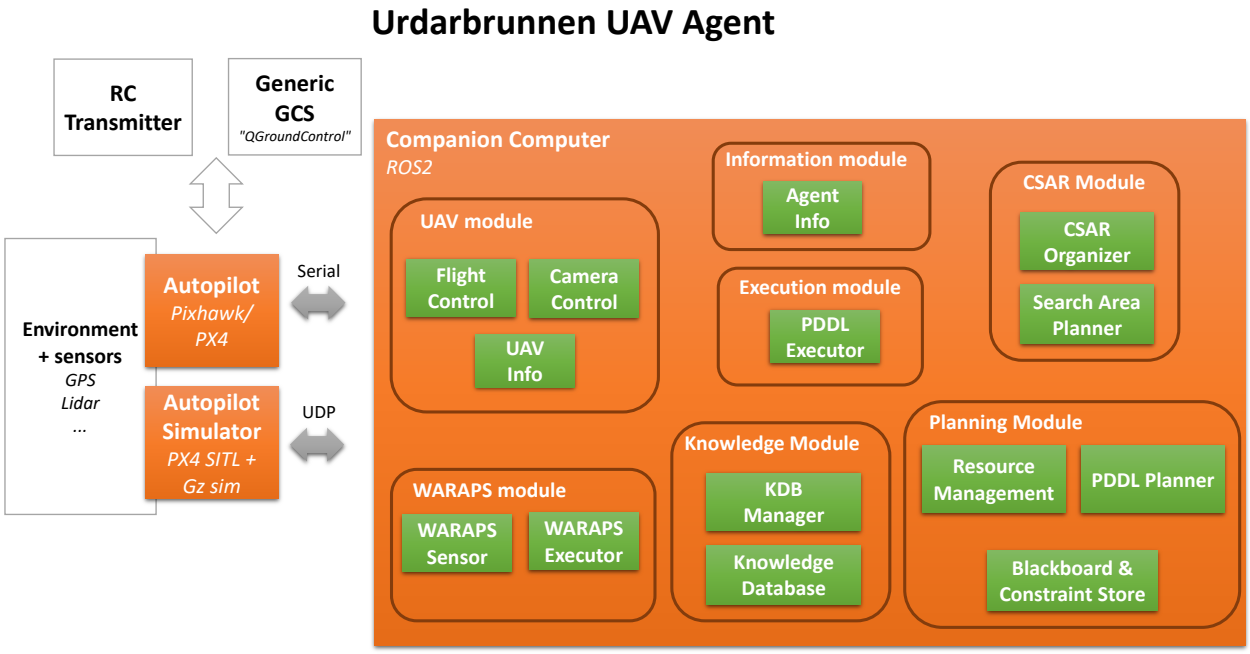

Fig. 3. Overarching view of the agent architecture, including hardware, (optional) software modules and interfaces.

when using a bottom-up initiative approach. When using a top-down approach the top-level planner is responsible for coordinating constraints between agents. However when performing a CSAR mission, the Blackboard can be used to communicate findings between agents so that agents can abort the search and return home when the missing persons are found.

CSAR Module: This is a mission specific module. It contains the CSAR Organizer that is capable of putting together a plan for a CSAR mission and executing it. It knows what type of resources are needed and how to put everything together. It is also responsible for monitoring the execution and re-planning from the current state in case something goes wrong. The first iteration of this module use a centralized top-down approach where one agent is responsible for deriving a detailed mission plan. In future iterations, we will compare different levels of agent initiative and task abstractions when solving the same type of missions.

The CSAR Module also contains the Search Area Planner, which is a mission specific planner. This planner support the CSAR Organizer by dividing the search area into sub areas that can be searched efficiently by a single UAV.

Knowledge Module: The knowledge module deals with the synchronization and combination of information at different levels of abstraction. It is responsible for storing and making accessible all types of knowledge ranging from high level semantic data down to low-level sensor data, such as stored images, LiDAR measurements, or behavioral models informed by doctrine. This information can be shared between agents in order to build up situational awareness. One framework for knowledge representation is the SymbiCloud framework described in [26].

WARA-PS Module: The WARA-PS module contains components that are used to make the UAV agents compatible with the JSON API-specification for WARA-PS [6]. This module makes it possible to provide sensor information (physical state and camera stream) to the WARA-PS core system (Sensor Agent in [6]), but also the possibility to execute tasks (Direct Execution Agent in [6]) or even plan and delegate missions.

D. Simulation

This section describes the simulation environment, which is used to perform initial test and validation of the ROS2 software modules.

PX4 SITL: With PX4 software-in-the-loop (SITL), it is possible to simulate the entire PX4 system, including the flight controller, sensors, and actuators, without performing real flights. This enables the possibility to test different configurations and settings of the autopilot prior to actual flights. PX4 SITL works by running the PX4 firmware on a virtual machine, which in turn communicates with a simulated environment through a network interface. The simulated environment can be a 3D robotics simulator, such as Gazebo or jMAVSim, or a custom simulator created by the user.

Gazebo: Gazebo [27] is an open-source 3D robotics simulator that allows for simulation of unmanned vehicles in various environments and scenarios. Gazebo provides a realistic simulation environment that can simulate the physics

of a UAV in flight. To simulate a UAV in Gazebo, one must first create a model of the UAV using so called simulation description format (SDF) files. These files define the physical properties of the UAV, such as its mass, size, shape, and aerodynamic properties. It is also possible to add and define sensors and other components of the UAV, such as camera, GNSS receiver, IMU, LiDAR, and motors. Once the model of the vehicle is created, it can be imported into Gazebo and placed in a virtual environment.

VI. CSAR Mission Example

In this section we explain the messages sent within the UAS when performing a CSAR mission. The message sequence is illustrated in Figure 4. The system requires an external Tasking Authority that could be a human using a mission system. The Tasking Authority is responsible for populating the Resource Management component with available UAV agents and tasking the CSAR Organizer with the mission. The CSAR Organizer starts with allocating resources by asking the Resource Management for all agents with flying and camera capabilities. The CSAR Organizer then asks each agent for information regarding its search capabilities. This information is then forwarded to the Search Area Planner that decomposes the search area into smaller strips, each of which can be searched efficiently by a single agent. Using this decomposition, the CSAR Organizer puts together a PDDL problem file and sends this together with the CSAR PDDL domain file to PlanSys2 and receives a plan. The CSAR Organizer releases all agents that it will not make use of to the Resource Management and provides timing information for when the allocated agents will be used. It then starts executing the mission. Finally the CSAR Organizer reports that the mission is complete back to the Tasking Authority.

VII. Conclusions and Future Work

In this paper, we have presented ongoing work from the first phase of the Urdarbrunnen project which aim is to iteratively develop an operator-assisted AI-enabled mission system for basic autonomous functions. In this first iteration, we have focused on finding relevant open-source, readily available components to swiftly and effectively, obtain an initial partial CSAR mission capability. This capability will be iteratively developed and it is now designed to be capable of performing a multi-agent search operation, in a complex and dynamic environment, while providing human machine interaction support for remote operators. Future work will involve simulations and real-world experiments in order to gain experimental results and to demonstrate the efficacy of the proposed system. As the aim the project is to develop this system iteratively, the capability of the system will be extended to incorporate additional tasks for other scenarios in future work. These scenarios may involve the integration of additional sensors and technologies, such as thermal cameras and LiDAR, to enhance the system's ability to detect objects and provide more accurate information to human operators. The system could also be extended to include more decentralized planning, with agents capable of planning a task of higher level of abstraction by themselves. Another extension is to include machine learning models that enable the UAV agents to learn from experience. Additionally, further research could focus on the integration and evaluation of communication protocols to enable seamless collaboration and coordination among the drones and human operators.

References

[1] FOI, "Totalförsvarets utmaningar på sikt kräver omfattande satsning på forskning idag." https://www.foi.se/nyheter-och-press/nyheter/2023-03-02-totalforsvarets-utmaningar-pa-sikt-kraver-omfattande-satsning-pa-forskning-idag.html, 2023. Last accessed March 2023.

[2] European Union Aviation Safety Agency, "Commission implementing regulation (EU) 2021/664 of 22 April 2021 on a regulatory framework for the U-space." https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32021R0664&from=EN, 2021. Last accessed March 2023.

[3] Saab AB, "Saab Purpose & Values." https://www.saab.com/about/company-in-brief/purpose-and-values, 2023. Last accessed March 2023.

[4] Saab, "Fighter systems." https://www.saab.com/products/air/fighter-systems. Last accessed March 2023.

[5] P. Doherty, J. Kvarnström, M. Wzorek, P. Rudol, F. Heintz, and G. Conte, "HDRC3 - A distributed hybrid deliberative/reactive architecture for unmanned aircraft systems," *Handbook of Unmanned Aerial Vehicles*, pp. 849–952, 2014.

[6] O. Andersson, P. Doherty, M. Lager, J.-O. Lindh, L. Persson, E. A. Topp, J. Tordenlid, and B. Wahlberg, "WARA-PS: a research arena for public safety demonstrations and autonomous collaborative rescue robotics experimentation," *Autonomous Intelligent Systems*, vol. 1, pp. 1–31, 2021.

[7] United States Department of Defense, "Unmanned Aircraft Systems (UAS) Roadmap, 2005–2030," 2005.

[8] Swedish Armed Forces, "Överbefälhavarens råd avseende förmågeutveckling (FM2022-19979:13)," 2022. In Swedish only.

[9] J. Gertler, "U.S. unmanned aerial systems," United States Congressional Research Service, 2012.

[10] A. C. Watts, V. G. Ambrosia, and E. A. Hinkley, "Unmanned aircraft systems in remote sensing and scientific research: Classification and considerations of use," *Remote sensing*, vol. 4, no. 6, pp. 1671–1692, 2012.

[11] Vinnova, "Autonomous search system, AuSSys." https://www.vinnova.se/p/autonomous-search-system-aussys/, 2022. Last accessed March 2023.

[12] J. Cacace, A. Finzi, V. Lippiello, M. Furci, N. Mimmo, and L. Marconi, "A control architecture for multiple drones operated via multimodal interaction in search & rescue mission," in *Proceedings of the 2016 IEEE International Symposium on Safety, Security, and Rescue Robotics (SSRR)*, pp. 233–239, IEEE, 2016.

[13] A. Atyabi, S. MahmoudZadeh, and S. Nefti-Meziani, "Current advancements on autonomous mission planning and management systems: An AUV and UAV perspective," *Annual Reviews in Control*, vol. 46, pp. 196–215, 2018.

[14] United States Air Force, "Combat Search and Rescue: Air Force Doctrine Document (AFDD) 2-1.6," 1998.

[15] United States Air Force, "Air Force Task List (AFTL) Air Force Doctrine Document (AFDD) 1-1," 1998.

[16] M. Ghallab, A. Howe, C. Knoblock, D. McDermott, A. Ram, M. Veloso, D. Weld, and D. Wilkins, "PDDL – the planning domain definition language (CVC TR-98-003/DCS TR-1165)," tech. rep., Yale Center for Computational Vision and Control, 1998.

[17] M. Fox and D. Long, "PDDL2.1: An extension to PDDL for expressing temporal planning domains," *Journal of Artificial Intelligence Research (JAIR)*, vol. 20, pp. 61–124, 2003.

[18] M. Colledanchise and P. Ögren, *Behavior trees in robotics and AI: An introduction*. CRC Press, 2018.

[19] S. Macenski, T. Foote, B. Gerkey, C. Lalancette, and W. Woodall, "Robot Operating System 2: Design, architecture, and uses in the wild," *Science Robotics*, vol. 7, no. 66, pp. 1–12, 2022.

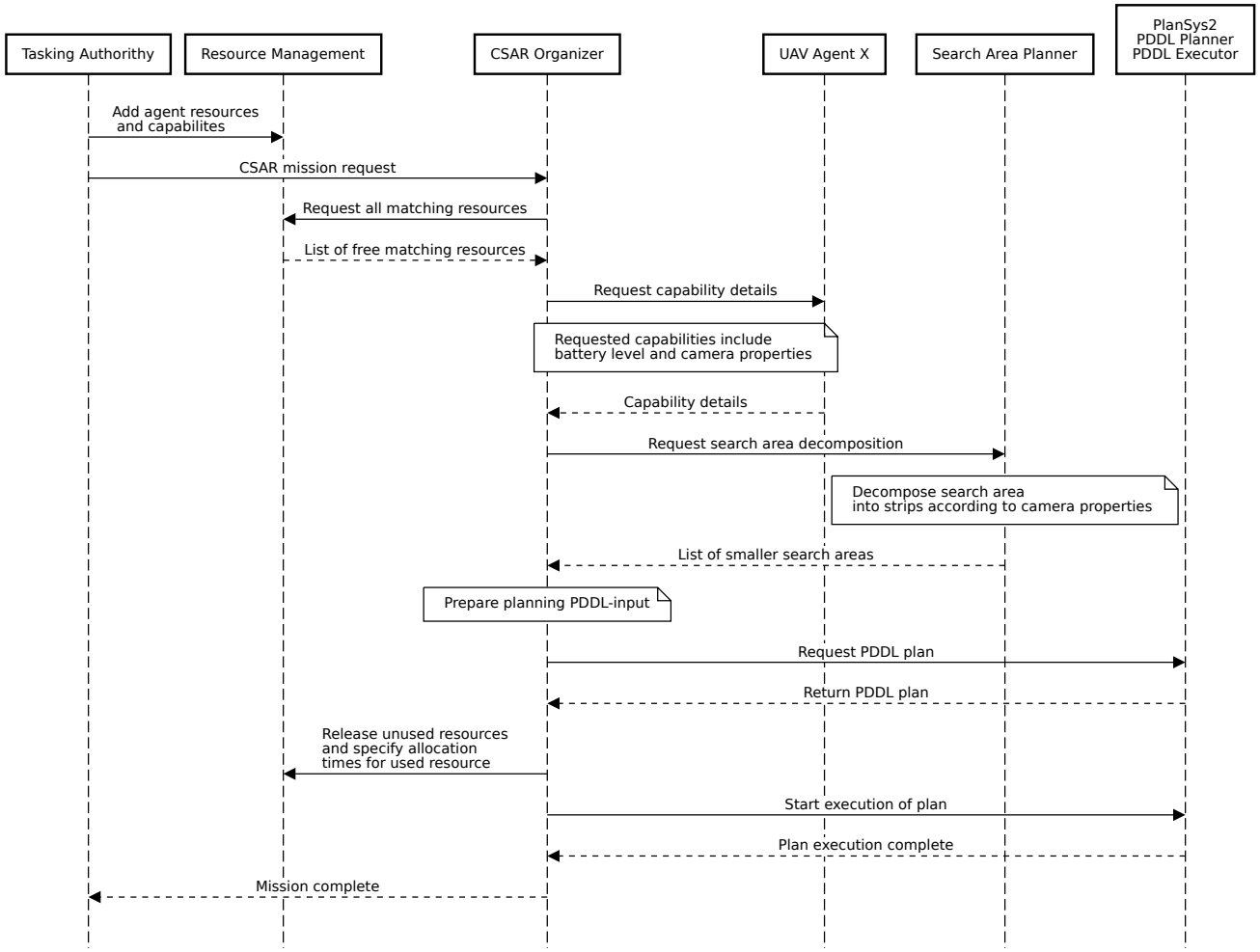

Fig. 4. Message sequence chart for a CSAR mission.

[20] G. Pardo-Castellote, "OMG data-distribution service: Architectural overview," in *Proceedings of the 23rd IEEE International Conference on Distributed Computing Systems (ICDCS)*, pp. 200–206, IEEE, 2003.

[21] L. Meier, D. Honegger, and M. Pollefeys, "PX4: A node-based multithreaded open source robotics framework for deeply embedded platforms," in *Proceedings of the 2015 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 6235–6240, IEEE, 2015.

[22] eProsima, "Micro XRCE-DDS agent." `https://github.com/eProsima/Micro-XRCE-DDS-Agent`, 2023. Last accessed March 2023.

[23] F. Martín, J. G. Clavero, V. Matellán, and F. J. Rodríguez, "PlanSys2: A planning system framework for ROS2," in *Proceedings of the 2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 9742–9749, IEEE, 2021.

[24] A. Coles, A. Coles, M. Fox, and D. Long, "Forward-chaining partial-order planning," in *Proceedings of the Twentieth International Conference on Automated Planning and Scheduling (ICAPS)*, vol. 20, pp. 42–49, 2010.

[25] P. Eyerich, R. Mattmüller, and G. Röger, "Using the context-enhanced additive heuristic for temporal and numeric planning," in *Towards Service Robots for Everyday Environments: Recent Advances in Designing Service Robots for Complex Tasks in Everyday Environments*, pp. 49–64, Springer, 2012.

[26] P. Doherty, C. Berger, P. Rudol, and M. Wzorek, "Hastily formed knowledge networks and distributed situation awareness for collaborative robotics," *Autonomous Intelligent Systems*, vol. 1, pp. 1–29, 2021.

[27] Open Robotics, "Gazebo." `https://gazebosim.org/home`. Last accessed March 2023.

Paper session 2: AI for Security and User Management

Evaluation of Defense Methods Against the One-Pixel Attack on Deep Neural Networks

Victor Arvidsson[1], Ahmad Al-Mashahedi[2] and Martin Boldt[3]

***Abstract*— The one-pixel attack is an image attack method for creating adversarial instances with minimal perturbations, i.e., pixel modification. The attack method makes the adversarial instances difficult to detect as it only manipulates a single pixel in the image. In this paper, we study four different defense approaches against adversarial attacks, and more specifically the one-pixel attack, over three different models. The defense methods used are: *data augmentation*, *spatial smoothing*, and *Gaussian data augmentation* used during both training and testing. The empirical experiments involve the following three models: all convolutional network (CNN), network in network (NiN), and the convolutional neural network VGG16.**

Experiments were executed and the results show that Gaussian data augmentation performs quite poorly when applied during the prediction phase. When used during the training phase, we see a reduction in the number of instances that could be perturbed by the NiN model. However, the CNN model shows an overall significantly worse performance compared to no defense technique. Spatial smoothing shows an ability to reduce the effectiveness of the one-pixel attack, and it is on average able to defend against half of the adversarial examples. Data augmentation also shows promising results, reducing the number of successfully perturbed images for both the CNN and NiN models. However, data augmentation leads to slightly worse overall model performance for the NiN and VGG16 models. Interestingly, it significantly improves the performance for the CNN model.

We conclude that the most suitable defense is dependent on the model used. For the CNN model, our results indicate that a combination of data augmentation and spatial smoothing is a suitable defense setup. For the NiN and VGG16 models, a combination of Gaussian data augmentation together with spatial smoothing is more promising. Finally, the experiments indicate that applying Gaussian noise during the prediction phase is not a workable defense against the one-pixel attack.

I. Introduction

Machine learning (ML), which is an important subarea of artificial intelligence (AI), has become both increasingly important and relevant during the last decades due to, for instance, its widespread use in critical applications. An important field within ML is *adversarial machine learning*, which is the study on how ML models can be attacked or deceived by antagonistic actors, i.e., adversaries [1]. Adversarial ML involves both the development of various attack methods against ML methods, as well as the development of defense methods against such attacks. These defense methods aim to improve the robustness of ML models [2].

Adversarial ML attacks are aimed either at the training data (e.g., data-poisoning attacks), the ML model's parameters, or the inputs during inference while using the ML model (e.g., adversarial input attacks) [3]. In general, such attacks can result in ML models that make incorrect predictions, which can result in serious consequences depending on the application, e.g., healthcare or autonomous vehicles.

In essence, this paper presents experimental results that evaluate defense methods against a particular adversarial ML attack referred to as the one-pixel attack [4]. We use the following three different models to evaluate the performance of the defenses: an all convolutional network (CNN), a Network in Network (NiN), and the convolutional neural network VGG16. The motivation for the choice of these three particular models is that those are the models used in the original paper describing the one-pixel attack [4]. In the experiments, two different defense methods are applied during the training phase of the models, while two different defense method are applied during the prediction phase.

The remainder of this paper is outlined as follows. Next, in Section II, the background is described, which includes the one-pixel attack and the applied defense methods. Then follows the related work in Section III, and then the methods description in Section IV. The results are presented in Section V followed by the analysis and discussion in Section VI. Finally, conclusions and future work are described in Section VII.

II. Background

In the background section, we describe the one-pixel attack as well as the defense methods evaluated in this study.

A. One-Pixel Attack

The one-pixel attack is an iterative semi-black-box attack that targets image recognition models. The main idea is to only perturb, i.e., modify, the value of one single pixel in an image to make the model miss-classify the whole image. This makes the attack harder to detect for humans when used on larger images. It also demonstrates that current models are not robust enough to ignore small adversarial perturbations [4].

B. Differential Evolution

In order to execute the one-pixel attack, differential evolution is used. Differential evolution is an evolutionary algorithm that minimizes a function value by creating candidate

*This work was not supported by any organization

[1]V. Arvidsson is a master's student at the Department of Computer Science, Blekinge Institute of Technology, Karlskrona, Sweden h.victor.arvidsson at gmail.com

[2]A. Al-Mashahedi is a master's student at the Department of Computer Science, Blekinge Institute of Technology, Karlskrona, Sweden ahmad.sebbah at gmail.com

[3]M. Boldt is a researcher at the Department of Computer Science, Blekinge Institute of Technology, Karlskrona, Sweden martin.boldt at bth.se

solution vectors and evaluating their fitness on a function. The method has three different parameters: the population size, $NP \geq 4$, the crossover probability, $CR \in [0, 1]$, and the differential weight, $F \in [0, 2]$. For each generation, each candidate solution, x, in that generation is mutated using three other distinct candidates solutions, a, b, and c, in combination with the F and CR parameters. A mutation vector is calculated according to Equation 1:

$$a + F \cdot (b - c). \tag{1}$$

For each dimension index, j in the mutation vector, a uniform random number $n \in [0, 1]$ is generated. A random integer index $R \in [1, NP]$ is also generated. If $n < CR$ or $j = R$, the value from the mutation vector at index j is selected for the new candidate, otherwise the value of x at index j is selected. The new candidate solution is compared to the old candidate x, and if it has a better performance it is added to the population as a replacement for x in the new generation. This continues for a specified number of generations, or until a stop criteria is met [5].

C. Defenses

For this report, we will evaluate three different defense methods. These are data augmentation, spatial smoothing, and Gaussian data augmentation. Some of these can be combined, which is discussed further in Section IV.

1) Data Augmentation: Data augmentation is used to increase the size of a dataset by adding slightly altered versions of the existing data points. This can increase the robustness of the trained model, as well as reduce overfitting. For images, this is usually done by applying an affine transformation, i.e., a linear transformation with translation, to the images. This may include rotating, translating, shearing, mirroring, and zooming the images [6].

2) Spatial Smoothing: Spatial smoothing is a feature squeezing method that reduces noise in the image by blurring it. There are two types of spatial smoothing: local and non-local. This report will focus on the local variant. Local spatial smoothing works by using information from nearby pixels to smooth each pixel. A sliding window passes over the image and updates each pixel according to a weighted kernel. This can perform different types of smoothing, such as Gaussian or median smoothing [7].

3) Gaussian Data Augmentation: The idea behind Gaussian data augmentation is to apply Gaussian noise to the inputs to the model. Gaussian data augmentation can be applied both during the training phase and the prediction phase. If used during the training phase, the training data is perturbed with Gaussian noise, similar to how regular data augmentation works. The dataset can either be augmented with the new samples, or replaced by them. If used during the prediction phase, the Gaussian noise is added to the input before it is passed into the model for prediction [8].

III. Related Works

In this section, the related work is summarized and the identified research gap is stated.

First, Bracamonte *et al.* proposed a novel approach, OPA2D, which is an extended one-pixel attack approach that aims to deceive humans and DNNs [9]. Further, they proposed to limit the attacked pixel RGB range in order to make it harder to detect by human vision. They show that an already attacked image, if attacked once again, tends to return to its original label. The results they achieve were good with detection rates up to 100% and defense rates between 93%-95%.

Husnoo *et al.* suggest an approach that utilizes robust principle component analysis and accelerated proximal gradient to detect the attacked pixel and recover it from the image [10], thus creating a clean non-attacked image with no deterioration in image quality.

Chen *et al.* proposed a Patch Selection Denoiser (PSD) approach to remove potential attacking pixels from an image without changing a large number of the pixels in the image [11]. The proposed approach achieved a 98.6% defense rate against one-pixel attacks. However, it relies on patching images independent of whether an adversarial image was detected or not, which degrades images due to the use of the denoising model.

Bennamoun *et al.* proposed an adversarial detection network (ADnet) that can detect adversarial pixels in images for robotic systems [12]. The authors claimed that it works as a defense method by rejecting adversarial examples. The performance of their approach, in the context of detecting adversarial scenarios, was evaluated using three different datasets. In the evaluation they perturbed 50% of the images, using 1, 3, and 5-pixel attacks, in order to create adversarial examples for the attack scenarios. The results indicate that ADNet's efficacy in detecting adversarial N-pixel attacks across the three datasets were shown by a detection accuracy above 90% for all datasets and attack types.

Tso *et al.* proposed a three-stage noise elimination and reconstruction algorithm in which they remove N-attacked pixels and then reconstruct the image, while at the same time keeping its integrity [13]. Their approach is to construct a difference map to evaluate the difference between pixels, as well as an average map to correspond with it. If the difference between a pixel and its neighboring pixel is deemed too high, it is replaced by the value of its neighboring pixel. The approach can be seen as a pre-processing step, so no model re-training is needed. The experiment results they achieved reveal that the proposed algorithm provides a protection rate ranging between 90% to 92% against N-pixel attacks, for N values of 1, 3, 5, 10, and 15.

A. Identified Research Gap

Adversarial ML attacks have been shown to be rather effective in deceiving ML models, which has highlighted aspects regarding security and reliability within ML systems. Therefore, research in this area is essential to addressing the problems raised by adversarial ML attacks, and to mitigate the exploitation of such attacks by malicious actors. This motivates this study in which we evaluate different defense methods against the one-pixel attack, which is an attack that

Fig. 1. The result of applying the spatial smoothing and Gaussian noise filters on a sample image.

is difficult to detect due to its low degree of perturbation. Thus, the added value through this work is the evaluation of defenses against the one-pixel attack on the same three deep learning models on which the original attack was evaluated.

IV. Method

In this section we describe the method used, e.g., the dataset, models and their configurations, as well as the experimental setup. In essence, the experiments evaluate the suitability of defense methods against the one-pixel attack, during both the training and the prediction phase, using three different deep neural network models.

A. Dataset Used

For evaluation of the defense methods against the one-pixel attack, we use the CIFAR-10 dataset [14]. CIFAR-10 contains 60,000 images across 10 different classes. The dataset is provided as 50,000 training images and 10,000 test images. Each image is 32x32 pixels and each pixel has three color channels: red, green, and blue. Each channel has integer values in the range $[0, 255]$, normalized into the range $[0, 1]$. The motivation for choosing the CIFAR-10 dataset is two-fold, first that it is a commonly used dataset in applied ML vision research, and second that it was used in the original study presenting the one-pixel attack [4].

B. Models

To evaluate the defenses we use three different models. The first model is the all convolution network (CNN) [15]. This uses nine convolutional layers of different sizes. The second model is a Network-in-Network model (NiN) [16]. This also uses nine convolutional layers, but introduces pooling layers between every third layer. The third model is the VGG16 [17]. It is a convolutional model that uses 13 convolutional layers, five max pooling layers, and two fully connected layers. The structures of the networks can be seen in Tables I, II, and III. The three network models are identical to the models used in the original one-pixel attack paper [4].

TABLE I
Model Summary for All Convolutional Network (CNN).

Conv2D(filters=96, kernel_size=3, stride=1, activation=ReLU)
Conv2D(filters=96, kernel_size=3, stride=1, activation=ReLU)
Conv2D(filters=96, kernel_size=3, stride=2, activation=ReLU)
Conv2D(filters=192, kernel_size=3, stride=1, activation=ReLU)
Conv2D(filters=192, kernel_size=3, stride=1, activation=ReLU)
Dropout(0.3)
Conv2D(filters=192, kernel_size=3, stride=2, activation=ReLU)
Conv2D(filters=192, kernel_size=3, stride=2, activation=ReLU)
Conv2D(filters=192, kernel_size=1, stride=1, activation=ReLU)
Conv2D(filters=10, kernel_size=1, stride=1, activation=ReLU)
GlobalAveragePooling2D
Flatten
Softmax

TABLE II
Model Summary for Network in Network (NiN).

Conv2D(filters=192, kernel_size=5, stride=1, activation=ReLU)
Conv2D(filters=160, kernel_size=1, stride=1, activation=ReLU)
Conv2D(filters=96, kernel_size=1, stride=1, activation=ReLU)
MaxPooling2D(pool_size=3, stride=2)
Dropout(0.5)
Conv2D(filters=192, kernel_size=5, stride=1, activation=ReLU)
Conv2D(filters=192, kernel_size=5, stride=1, activation=ReLU)
Conv2D(filters=192, kernel_size=5, stride=1, activation=ReLU)
AveragePooling2D(pool_size=3, stride=2)
Dropout(0.5)
Conv2D(filters=192, kernel_size=3, stride=1, activation=ReLU)
Conv2D(filters=192, kernel_size=1, stride=1, activation=ReLU)
Conv2D(filters=10, kernel_size=1, stride=1, activation=ReLU)
GlobalAveragePooling2D
Flatten
Softmax

TABLE III
Model Summary for VGG16.

Conv2D(filters=64, kernel_size=3, stride=1, activation=ReLU)
Conv2D(filters=64, kernel_size=3, stride=1, activation=ReLU)
MaxPooling2D(pool_size=2, stride=2)
MaxPooling2D(pool_size=2, stride=2)
Conv2D(filters=128, kernel_size=3, stride=1, activation=ReLU)
Conv2D(filters=128, kernel_size=3, stride=1, activation=ReLU)
MaxPooling2D(pool_size=2, stride=2)
Conv2D(filters=256, kernel_size=3, stride=1, activation=ReLU)
Conv2D(filters=256, kernel_size=3, stride=1, activation=ReLU)
Conv2D(filters=256, kernel_size=3, stride=1, activation=ReLU)
MaxPooling2D(pool_size=2, stride=2)
Conv2D(filters=512, kernel_size=3, stride=1, activation=ReLU)
Conv2D(filters=512, kernel_size=3, stride=1, activation=ReLU)
Conv2D(filters=512, kernel_size=3, stride=1, activation=ReLU)
MaxPooling2D(pool_size=2, stride=2)
Conv2D(filters=512, kernel_size=3, stride=1, activation=ReLU)
Conv2D(filters=512, kernel_size=3, stride=1, activation=ReLU)
Conv2D(filters=512, kernel_size=3, stride=1, activation=ReLU)
MaxPooling2D(pool_size=2, stride=2)
Flatten
Dense(2048, activation=ReLU)
Dense(2048, activation=ReLU)
Dense(10, activation=Softmax)

C. Attack

The one-pixel attack uses differential evolution, and in this study we choose a population size of 400, a crossover probability of 1, and a differential weight that is uniformly randomized between 0.5 and 1 for each generation. These parameters where chosen based on the parameters in the original paper. Using a randomized differential weight can speed up convergence. Each candidate solution is an array of the following five values:

1) X coordinate for the perturbed pixel (between 0-31).
2) Y coordinate for the perturbed pixel (between 0-31).
3) Value for the red color channel (between 0-1).

TABLE IV
THE NINE DIFFERENT MODELS, BASED ON DATASET AND NETWORK ARCHITECTURES, THAT WERE EVALUATED.

		Models		
		CNN	**NiN**	**VGG16**
Dataset	**Original**	CNN_{orig}	NiN_{orig}	$VGG16_{orig}$
	Augmented	CNN_{aug}	NiN_{aug}	$VGG16_{aug}$
	Gaussian	CNN_{gau}	NiN_{gau}	$VGG16_{gau}$

4) Value for the green color channel (between 0-1).
5) Value for the blue color channel (between 0-1).

To evaluate each candidate solution, the image is perturbed with the candidate pixel, and the class is predicted by the classifier. We only perform untargeted attacks, and the goal is therefore to minimize the certainty of the model for the true label. The differential evolution runs for a total of 100 generation for each image. We include an early stop condition when the confidence for the true label is lower than 5%.

D. Defenses

We implement the four different types of defenses discussed in Section II. Two defenses applied during the training phase of the models, and two defenses applied during the testing phase.

1) Training-Phase Defenses: The first type of defenses are the ones that are applied during model training phase. These are the data augmentation and Gaussian data augmentation defenses, which are denoted CNN_{aug} and CNN_{gau} respectively for the CNN model. Both of these generate a new dataset with the augmented images, and for each of these a separate instance of each network type is trained. This results in a total of nine different models, as can be seen in Table IV. The applied augmentations for the data augmentation defense are: rotation up to 20° in each direction, width and height shift up to 20% of the image's width and height respectively, shearing with a maximum shear angle of 20°, zooming with a maximum range of 20% for both zoom-in and zoom-out, and finally mirroring along the vertical axis. For the Gaussian data augmentation, the noise is generated with a standard deviation of 0.05, and the samples were augmented at a ratio of 0.5, which means that the size of the dataset increases by 50%.

2) Testing-Phase Defenses: The testing-phase defenses are applied to each adversarial instance before the model performs the classification. Each of these defenses were applied separately to each instance. The spatial smoothing defense uses the median strategy for smoothing with a window size of 3. The Gaussian data augmentation applies Gaussian noise to the instance with a standard deviation of 0.05. Each testing-phase defense is tested separately and in combination with each training defense for each model.

E. Evaluation Metrics

To evaluate the performance of the models, we use both accuracy and Area Under the ROC Curve (AUC) score. For evaluation of the defense methods, we use normalized defense ratio, which is calculated according to Equation 2:

$$D_R = \frac{C_d}{C_p} \tag{2}$$

where C_d is the number of correctly classified instances after *both* the attack and defense methods were employed, while C_p is the number of correctly classified instances *before* the attack when only the defense was applied. To get the normalized defense ratio, we scale D_R by the accuracy of the model on the total attacked instances. Thus, the normalized defense ratio metric is calculated according to Equation 3:

$$\overline{D}_R = D_R \frac{C_p}{N} = \frac{C_d}{N} \tag{3}$$

where N is the total number of instances, in our case 1,000.

F. Experimental Setup

In the experimental evaluation of the models performance, the independent variable was the candidate models, i.e., CNN, NiN, and VGG16. The dependent variables were the accuracy and the AUC scores. For the experimental evaluation of the defense methods, the independent variables were the model candidates, the training-phase defense methods, and the testing-phase defense methods. The dependent variable was the defense ratio metric.

The three models that used the Gaussian augmentation defense (CNN_{gau}, NiN_{gau}, and $VGG16_{gau}$) and the three original models (CNN_{orig}, NiN_{orig}, and $VGG16_{orig}$) were trained during 100 epochs. Due to the computationally demanding proces to the generate and predict the image data, the models that used the augmented defense (CNN_{aug}, NiN_{aug}, and $VGG16_{aug}$) were trained for 70 epochs.

In total, 1,000 different images were randomly sampled from the test images in the CIFAR-10 dataset. The attack was performed individually on these 1,000 images for each of the three different network models included in the study and for each of the three different training datasets described in Section IV-D. This resulted in a total of 9,000 adversarial images that were included in the experiment.

The experiments were executed on a system with an Intel i7-6700K processor, Nvidia GTX 980 Ti graphics card and 16GB RAM.

V. RESULTS

The results are presented for the investigated models as well as the the one-pixel attack, including time measurements.

A. Model Performance

All models except the CNN_{gau} showed an AUC score above 0.9, with CNN_{aug} scoring best on both metrics with an AUC and accuracy of 0.982 and 0.85 respectively, see Fig. 2. Worst performance was associated with the Gaussian-augmented CNN model (CNN_{gau}) with an AUC and accuracy of 0.83 and 0.49 respectively. The remaining models show performance with AUC scores significantly above 0.9 and accuracy above 0.8 on average.

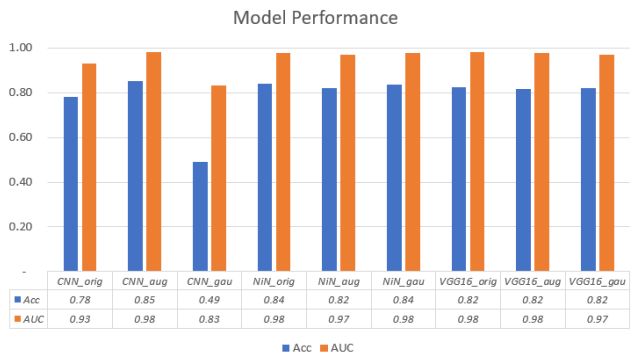

Fig. 2. Performance of the trained models on the entire CIFAR-10 test data set.

Fig. 3. Example of perturbed images, left image is from an attack on the CNN_{orig} model and the right one is from an attack on the $VGG16_{aug}$ model.

B. Results of the One-Pixel Attack

Two examples of attacked images and their associated classes are shown in Fig. 3. The prediction performance, one-pixel attack performance, and defense performance for three versions of the three models are shown in Fig. 4, 5, and 6. As an example, in Fig. 4, we can see that CNN_{orig} predicted the correct label for 801 images out of 1,000. The one-pixel attack was able to perturb 358 of these images so they were predicted with an erroneous class label. The spatial smoothing and Gaussian noise were able to defend against 203 and 160 out of the total 358 perturbed images respectively.

A complete overview of the results can be seen in the appendix.

It is worth noting that there were cases where the defense method caused the models to change their predictions from a correct label to an incorrect label, despite the attack not being successful. These cases are not presented since they are out of scope of this paper.

C. Execution Times

The process of running differential evolution for a single model takes approximately 14 seconds on average, for each unique image. With a total of nine models to evaluate the attack on results in a total execution time of 126 seconds, i.e., roughly two minutes, per unique image. For the 1,000 images the one-pixel attack process took around 35 hours to execute.

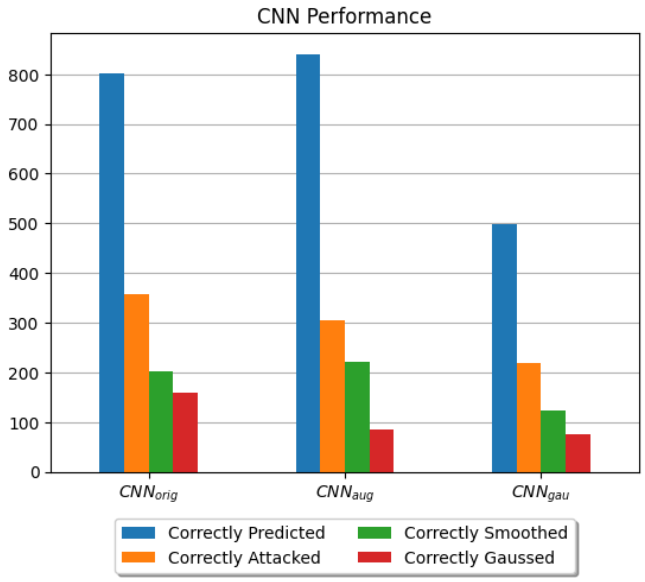

Fig. 4. Performance of the CNN models.

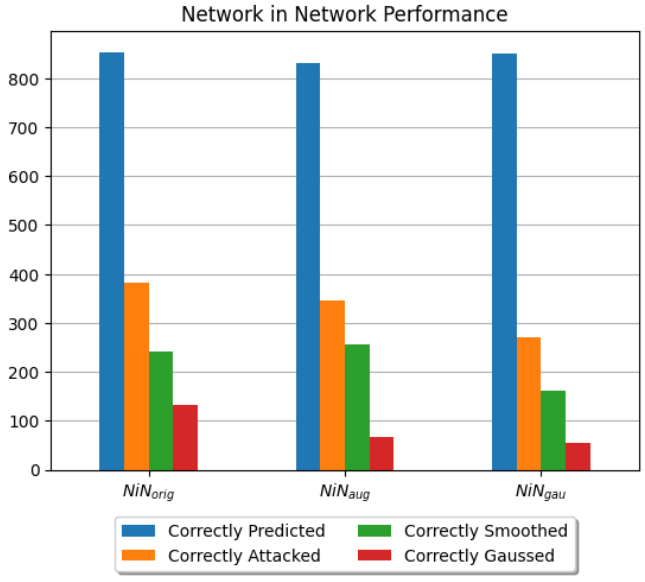

Fig. 5. Performance of the NiN models.

VI. Discussion

In this section, we will discuss the results for the original base models, for the different defenses, and for the one-pixel attack. The section is finished with a subsection that addresses validity threats connected to the experiments.

A. Base Models (NiN_{orig}, CNN_{orig}, and $VGG16_{orig}$)

The performance of the original models, CNN_{orig}, NiN_{orig}, and $VGG16_{orig}$, differed slightly according to the accuracy metric with 80%, 85% and 81% respectively, see Table V

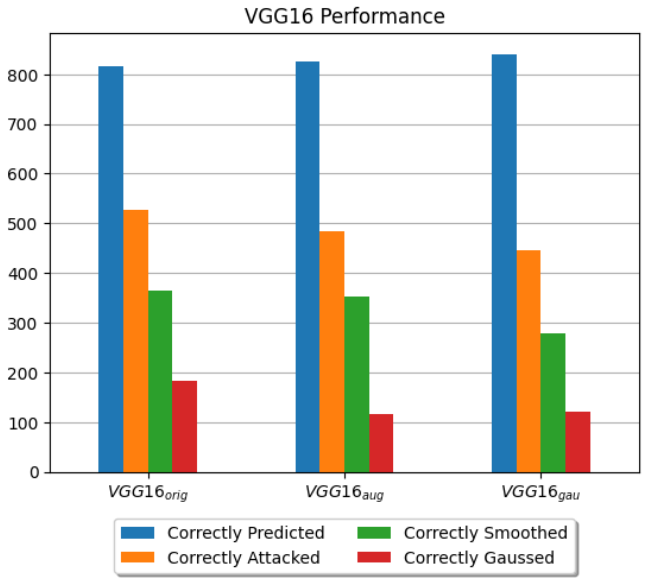

Fig. 6. Performance of the VGG16 models.

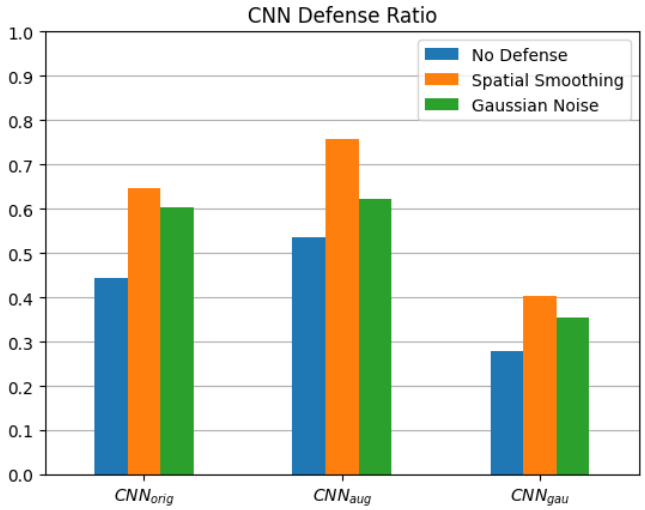

Fig. 7. defense ratio of the CNN models in relation to all 1,000 images.

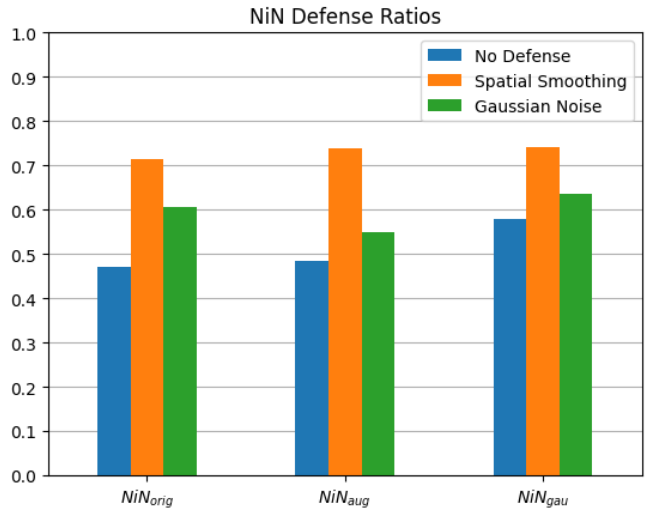

Fig. 8. defense ratio of the NiN models in relation to all 1,000 images.

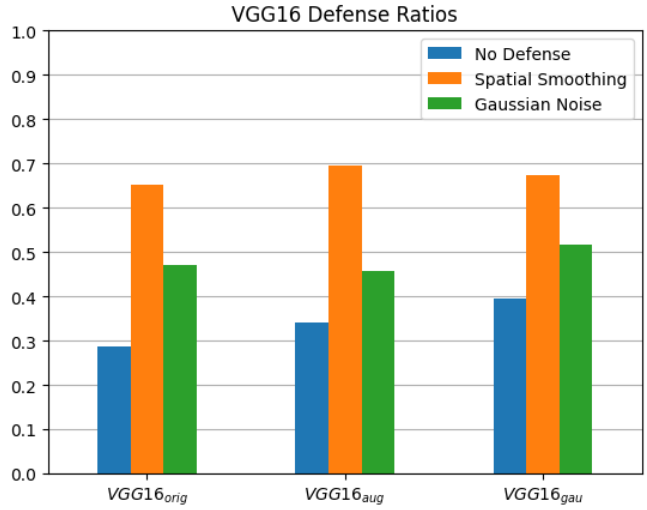

Fig. 9. defense ratio of the VGG16 models in relation to all 1,000 images.

in Appendix A. The AUC metric indicates similar results with measures of 0.93, 0.98, and 0.98 respectively. Overall, the NiN models achieved on average the highest prediction performance out of the three investigated model families. All investigated NiN models reached an accuracy of at least 83%, i.e., at least 830 correct predictions on the non-perturbed images, see Fig. 5. This is in line with what the authors of the model architecture have indicated as achieved state-of-the-art performance on the CIFAR-10 dataset [16]. The NiN model family was also the most difficult type of model to attack, since the defense ratios on average were higher than the other model families. The NiN_{orig} model achieved a defense ratio of 47.2% without applying any defense method, compared to the CNN_{orig} and $\text{VGG16}_{\text{orig}}$ models that achieved 44.3% and 28.8% respectively. This indicates that the NiN_{orig} model has a relatively higher inherent robustness compared to CNN_{orig} and $\text{VGG16}_{\text{orig}}$.

B. Defenses

1) Augmentation Defenses: Looking at Fig. 7, Fig. 8, Fig. 9, it is clear that in the majority of cases the defense ratio of the normal augmented and Gaussian augmented models is higher than the non-augmented models. The best augmentation method differs for the different model families. In the case of NiN and VGG16, Gaussian augmentation seems to be the better augmentation method, as the defense ratios in those cases are overall higher compared to their original counterparts. However, Gaussian augmentation performs significantly worse on the CNN models, as there is a clear performance deterioration compared to the other models. The normal augmentation method is associated with higher defense ratios on the CNN models, but slightly worse for

NiN and VGG16 models when compared to their Gaussian augmented counterparts. They do, however, perform better compared to their original non-augmented versions.

The experimental results indicate that the NiN models can use either of the two augmentation methods to improve the defense capabilities compared to having no defense at all. whilst the CNN and VGG16 models have increased defense capabilities for one of the augmentation methods and, depending on the model, worse for the other.

2) Spatial Smoothing: In all cases, the spatial smoothing defense was able to remove the visual representation of the attacked pixel through the blurring of the image. The problem arises when the model tries to predict on the smoothed image. It seems as if the models sometimes have difficulty in discerning the contents of the image, resulting in an incorrect prediction after the application of spatial smoothing. This seems to be specially true if the model's decision boundary for the particular class label is already uncertain from the start. There are, however, many cases where smoothing worked and the model was able to predict the correct label for the smoothed image. As seen in Fig. 4, Fig. 5 and Fig. 6 in most cases, spatial smoothing was able to correctly defend at least 50% of the attacked images, with certain models achieving up to 72% defended images. An interesting observation is that for the CNN_{orig} and CNN_{gau} models, spatial smoothing performed somewhat worse compared to the other models with spatial smoothing applied.

3) Gaussian Noise: Out of the two investigated defense methods that are applied during the prediction phase, the results indicate that Gaussian noise is the less effective defense. The results show a clear difference between the amount of perturbed images that each defense method successfully defended against, with spatial smoothing being the most suitable defense candidate. Even the Gaussian models that were trained on Gaussian noised data showed subpar performence when predicting the label on the noisy data. There are several possible reasons for this. First, the noising process could have had a too strong effect on the images, and in turn result in that the model could not properly recognize images due to the noise. Second, it could be due to the possibility that the Gaussian models were trained on too few Gaussian noised images and could therefore not capture enough variance in order to generalize properly.

Intuitively, the noised images that the Gaussian models trained on should increase the models robustness against attacks, and the ability to correctly predict correct labels on Gaussian noised images. However, it seems as there was an opposite effect as indicated by CNN_{gau} in Fig. 4. The Gaussian noise defense method performed better on models that were trained on Gaussian noised data compared to those that were not. Finally, it can be noted that adding Gaussian noise does not remove, or necessarily change, the perturbed pixel. Since the one-pixel attack relies on the model being heavily dependent on the value of the perturbed pixel, adding noise to the other pixels may not necessarily change the model's prediction.

C. One-Pixel Attack

The differential evolution algorithm was able to perturb 47% of the images that were correctly labeled for all of the models. This increases to 51% when considering only the original models. This is slightly worse than the results from the original paper, where they show a 68% success-rate for the attacks. One explanation for this difference could be that the original paper used the Kaggle version of the CIFAR-10 dataset, while we use the original CIFAR-10 dataset [4]. In the original paper the authors argued that the reason for this difference could be due to the fact that there is less noise in the original CIFAR-10 dataset, compared to the Kaggle version. This means that the models can achieve better training results, thus making the one-pixel attack less effective on the original dataset.

The combination of model and defense methods that resulted in the least number of attacked images is the same model that also showed the best prediction performance, i.e., the CNN_{aug} model using spatial smoothing. That particular model correctly predict the class label for 841 out of the 1,000 images. At the same time the one-pixel attack was successfully applied to 304 images, but the combination of normal augmentation and spatial smoothing managed to protect 221 out of these, i.e., leaving 83 images (or 8.3%) that were successfully attacked.

D. Experimental Validity Threats

As for any research method, the one chosen for this study is associated with a number of validity threats [18]. First there is an external validity threat as this study only investigates three different neural network model architectures. However, we attempt to address this by choosing the same models, including their network configuration, as the original paper describing the one-pixel attack. Another validity threat is due to the fact that the study only includes one dataset, and further, a sub-sample of images from that dataset. The reason for this is that the one-pixel attack and defense scenarios are quite computationally demanding, which limits the number of images that could be included in the study. Also, the chosen dataset is widely used in related research studies. Finally, regarding the sub-sampling a uniform random sampling was implemented.

VII. Conclusions and Future Work

The defense methods presented show potential for being effective against one-pixel attacks. The normal and Gaussian augmentation defense methods are more robust than their original counterparts, especially for the NiN and VGG16 models. Without applying any prediction defenses, Gaussian data augmentation provides the best defense for both VGG16 and NiN. Spatial smoothing seems to be the most effective defense method from a model-agnostic perspective, while Gaussian noise added during prediction shows slightly worse defense capability for all models. The ability to combine different defense methods shows an increased robustness, however, the most optimal combinations of the defenses vary

between the different network types. For the NiN, combining Gaussian data augmentation with spatial smoothing yielded the best results, while the CNN and VGG16 models worked best when combining normal augmentation and spatial smoothing. We can also conclude that the Gaussian noise defense worked best when combined with Gaussian augmentation, except for the CNN model. The combination of model and defense methods that showed best performance was CNN using a combination of normal augmentation and spatial smoothing, for which only 8.3% of the images were successfully attacked.

A. Future Work

An interesting idea for future research is to experiment with combining both augmentation methods with NiN and see if there is a big improvement in robustness over using one or the other. This can be further extended by training on smoothed images as well to see if there is an improvement when predicting on smoothed images.

Further research is also needed on the optimal parameters for the different defense methods and models. There is likely not one set of parameters that works for every situation, and a study of the optimal parameters for different situation might work as a basis when implementing these algorithms in new environments.

As mentioned in the results, we do not evaluate the effects of the defenses for non-adversarial images, and further research into any potential model degradation when using these defense methods is needed.

REFERENCES

[1] N. Papernot, P. McDaniel, S. Jha, M. Fredrikson, Z. B. Celik, and A. Swami, "The limitations of deep learning in adversarial settings," in *2016 IEEE European Symposium on Security and Privacy (EuroS&P)*, 2016, pp. 372–387.

[2] N. Carlini and D. Wagner, "Towards evaluating the robustness of neural networks," in *2017 IEEE Symposium on Security and Privacy (SP)*. Los Alamitos, CA, USA: IEEE Computer Society, may 2017, pp. 39–57. [Online]. Available: https://doi.ieeecomputersociety.org/10.1109/SP.2017.49

[3] I. Goodfellow, P. McDaniel, and N. Papernot, "Making machine learning robust against adversarial inputs," *Communications of the ACM*, vol. 61, 2018. [Online]. Available: https://dl.acm.org/doi/10.1145/3134599

[4] J. Su, D. V. Vargas, and K. Sakurai, "One pixel attack for fooling deep neural networks," *IEEE Transactions on Evolutionary Computation*, vol. 23, no. 5, pp. 828–841, 2019.

[5] R. Storn and K. Price, "Differential evolution – a simple and efficient heuristic for global optimization over continuous spaces," *Journal of Global Optimization*, vol. 11, no. 4, p. 341–359, 1997.

[6] L. Perez and J. Wang, "The effectiveness of data augmentation in image classification using deep learning," 2017. [Online]. Available: https://arxiv.org/abs/1712.04621

[7] W. Xu, D. Evans, and Y. Qi, "Feature squeezing: Detecting adversarial examples in deep neural networks," in *Proceedings 2018 Network and Distributed System Security Symposium*, vol. 1, 2018, pp. 289–303.

[8] J. F. R. Rochac, L. Liang, N. Zhang, and T. Oladunni, "A gaussian data augmentation technique on highly dimensional, limited labeled data for multiclass classification using deep learning," in *2019 Tenth International Conference on Intelligent Control and Information Processing (ICICIP)*, 2019, pp. 145–151.

[9] H.-Q. Nguyen-Son, T. P. Thao, S. Hidano, V. Bracamonte, S. Kiyomoto, and R. S. Yamaguchi, "Opa2d: One-pixel attack, detection, and defense in deep neural networks," in *2021 International Joint Conference on Neural Networks (IJCNN)*, 2021, pp. 1–10.

[10] M. A. Husnoo and A. Anwar, "Do not get fooled: Defense against the one-pixel attack to protect IoT-enabled deep learning systems," *Ad Hoc Networks*, vol. 122, p. 102627, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1570870521001499

[11] D. Chen, R. Xu, and B. Han, "Patch selection denoiser: An effective approach defending against one-pixel attacks," in *Neural Information Processing*, T. Gedeon, K. W. Wong, and M. Lee, Eds. Springer International Publishing, 2019, pp. 286–296.

[12] S. A. A. Shah, M. Beugre, N. Akhtar, M. Bennamoun, and L. Zhang, "Efficient detection of pixel-level adversarial attacks," in *2020 IEEE International Conference on Image Processing (ICIP)*, 2020, pp. 718–722.

[13] Z.-Y. Liu, P. S. Wang, S.-C. Hsiao, and R. Tso, "Defense against n-pixel attacks based on image reconstruction," in *Proceedings of the 8th International Workshop on Security in Blockchain and Cloud Computing*, ser. SBC '20. New York, NY, USA: Association for Computing Machinery, 2020, p. 3–7. [Online]. Available: https://doi.org/10.1145/3384942.3406867

[14] A. Krizhevsky, "Learning multiple layers of features from tiny images," 2009.

[15] J. T. Springenberg, A. Dosovitskiy, T. Brox, and M. Riedmiller, "Striving for simplicity: The all convolutional net," 2014. [Online]. Available: https://arxiv.org/abs/1412.6806

[16] M. Lin, Q. Chen, and S. Yan, "Network in network," 2013. [Online]. Available: https://arxiv.org/abs/1312.4400

[17] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014. [Online]. Available: https://arxiv.org/abs/1409.1556

[18] C. Wohlin, P. Runeson, M. Höst, M. C. Ohlsson, B. Regnell, and A. Wesslèn, *Experimentation in Software Engineering*. Heidelberg: Springer Berlin, 2012.

TABLE V
THE COMPLETE DEFENSE RESULTS FROM THE EXPERIMENTS.

		N	C_p	C_A	C_S	C_G	D_{R_O}	D_{R_S}	D_{R_G}	$\overline{D}_{R_O}$	$\overline{D}_{R_S}$	$\overline{D}_{R_G}$	t
Model	**CNN$_{\text{Orig}}$**	1,000	801	358	203	160	55.3%	80.6%	75.3%	44.3%	64.6%	60.3%	5.7
	CNN$_{\text{Aug}}$	1,000	841	304	221	85	63.9%	90.1%	74.0%	53.7%	75.8%	62.2%	12.3
	CNN$_{\text{Gau}}$	1,000	498	219	123	76	56.0%	80.7%	71.3%	27.9%	40.2%	35.5%	7.9
	NiN$_{\text{Orig}}$	1,000	854	382	242	133	55.3%	83.6%	70.8%	47.2%	71.4%	60.5%	10.7
	Nin$_{\text{Aug}}$	1,000	830	346	255	66	58.3%	89.0%	66.3%	48.4%	73.9%	55.0%	14.4
	NiN$_{\text{Gau}}$	1,000	850	270	161	55	68.2%	87.2%	74.7%	58.0%	74.1%	63.5%	12.5
	VGG16$_{\text{Orig}}$	1,000	815	527	364	182	35.3%	80.0%	57.7%	28.8%	65.2%	47.0%	8.5
	VGG16$_{\text{Aug}}$	1,000	825	484	353	116	41.3%	84.1%	55.4%	34.1%	69.4%	45.7%	16.4
	VGG16$_{\text{Gau}}$	1,000	841	446	279	121	47.0%	80.1%	61.4%	39.5%	67.4%	51.6%	11.0

APPENDIX

A. Result Table

In Table V we present the complete results from our experiments.

1) Variable explanation:

- N: The total number of instances tested.
- C_p: The total number of correctly predicted instances.
- C_A: The total number of correctly attacked instances from the correctly predicted instances.
- C_S: The total number of correctly defended instances with spatial smoothing from the correctly attacked instances.
- C_G: The total number of correctly defended instances with Gaussian noise from the correctly attacked instances.
- D_{R_O}: The defense ratio without applying any defense. Calculated as:
$$\frac{C_p - C_A}{C_p} = 1 - \frac{C_A}{C_p}.$$
- D_{R_S}: The defense ratio after applying spatial smoothing. Calculated as:
$$\frac{C_p - (C_A - C_S)}{C_p} = 1 - \frac{C_A - C_S}{C_p}.$$
- D_{R_G}: The defense ratio with after applying Gaussian noise. Calculated as:
$$\frac{C_p - (C_A - C_G)}{C_p} = 1 - \frac{C_A - C_G}{C_p}.$$
- $\overline{D}_{R_O}$: The normalized defense ratio without applying any defense. Calculated as:
$$D_{R_O} \cdot \frac{C_p}{N} = \frac{C_p - C_A}{N}.$$
- $\overline{D}_{R_S}$: The normalized defense ratio after applying spatial smoothing. Calculated as:
$$D_{R_S} \cdot \frac{C_p}{N} = \frac{C_p - C_A + C_S}{N}.$$
- $\overline{D}_{R_G}$: The normalized defense ratio after applying spatial smoothing. Calculated as
$$D_{R_G} \cdot \frac{C_p}{N} = \frac{C_p - C_A + C_G}{N}.$$
- t: The average time taken to perform the attack for the successfully attacked instances, measured in seconds.

Can the use of privacy-enhancing technologies enable federated learning for health data applications in a Swedish regulatory context?*

Rickard Brännvall[†], Helena Linge[†‡], Johan Östman[‡]

***Abstract*— A recent report by the Swedish Authority for Privacy Protection (IMY) evaluates the potential of jointly training and exchanging machine learning models between two healthcare providers. In relation to the privacy problems identified therein, this article explores the trade-off between utility and privacy when using privacy-enhancing technologies (PETs) in combination with federated learning. Results are reported from numerical experiments with standard text-book machine learning models under both differential privacy (DP) and Fully Homomorphic Encryption (FHE). The results indicate that FHE is a promising approach for privacy-preserving federated learning, with the CKKS scheme being more favorable in terms of computational performance due to its support of SIMD operations and compact representation of encrypted vectors. The results for DP are more inconclusive. The article briefly discusses the current regulatory context and aspects that lawmakers may consider to enable an AI leap in Swedish healthcare while maintaining data protection.**

I. Introduction

Recent advances in artificial intelligence (AI) have shown great promise in improving diagnosis, treatment, personalized medicine [1] and disease prevention by predictions [2]. Machine learning algorithms can analyze vast amounts of medical data, such as patient records, imaging scans, and genetic information, to identify patterns and make predictions about the likelihood of diseases and the effectiveness of treatments. Additionally, computer vision algorithms can analyze medical images, e.g. X-rays and MRI scans, detect abnormalities, and provide decision support in diagnosing disease.

However, the use of AI in medical applications raises concerns about privacy, as it involves the processing of sensitive personal information protected by privacy laws and regulations, such as the General Data Protection Regulation (GDPR) in the EU and the Health Insurance Portability and Accountability Act (HIPAA) in the USA, as well as country specific patient data regulations. To facilitate the sharing of personal health information between healthcare providers and digital health services, adequate privacy protection is essential. Full anonymization (de-identification) is often not possible as it impairs full utility of the data [3]. Therefore, several alternative approaches have been proposed, including cryptographic techniques, differential privacy, synthetic healthcare data generation, federated learning, and pseudonymization [4].

One technology that shows great potential in privacy preservation is fully homomorphic encryption (FHE) [5]. It makes computation on encrypted data possible which enables privacy-by-design cloud-based services. Federated learning allows multiple parties to collaboratively train a machine learning model without exchanging actual data. However, all comprehensive solutions must have a solid foundation in conventional security technology, policies, and procedures.

There is often a utility versus privacy trade-off when using privacy-enhancing technologies (PETs). However, for medical applications, a decreased utility translates into suboptimal data use, and loss of adequacy with regard to results and outcomes. If utility loss is allowed prolonged suffering and possibly even death may result. **How can the application of advanced privacy-enhancing measures in federated learning maintain a preserved privacy without the undue compromise of utility?** We limit our exploration of this question to two privacy problems: 1) that the final model parameters can potentially disclose personal data, and 2) that the sharing of model updates in the federated learning process can potentially disclose sensitive information from the respective parties' data sets.

Motivation: Regulators are currently investigating how PETs can unlock the potential of data-driven applications. We here explore in practice how these technologies can enable an AI leap in Swedish Healthcare already within the current privacy legislation, as well as identify questions that lawmakers may consider to achieve harmony with developments and demands.

Contribution: We discuss approaches to applying PETs to improve data protection in federated learning. We compare two quantum computer resilient FHE schemes and conclude that one has an advantage in terms of computational performance. Our experiments indicate that FHE is both feasible and favorable, as it preserves utility while adhering to data use minimization and purpose limitations. We also conduct numerical experiments with differential privacy (DP), which confirm the view that it, in its strictest form, may have significant utility degradation for trained models.

Outline: The next section provides Background to this work, providing both technical details on the advanced PETs and a summary of recent regulatory developments in Sweden. The Methods section explains the proposed approach. It describes the setup for the numerical experimentation, from which Results then are presented in the next section. The Discussion section presents an analysis of the pros and cons of the proposed alternative use cases, both in relation to the results from the numerical experiments and in relation to previously published work. The article then concludes with some recommendations.

II. Background

A. Machine learning context

Machine learning is a subfield of artificial intelligence that involves training algorithms to learn patterns in data without being explicitly programmed. Deep learning refers to algorithms

*This work was supported by Vinnova grant 2022-02668 (HEIDA).

[†]RISE Research Institutes of Sweden.
[‡]AI Sweden.

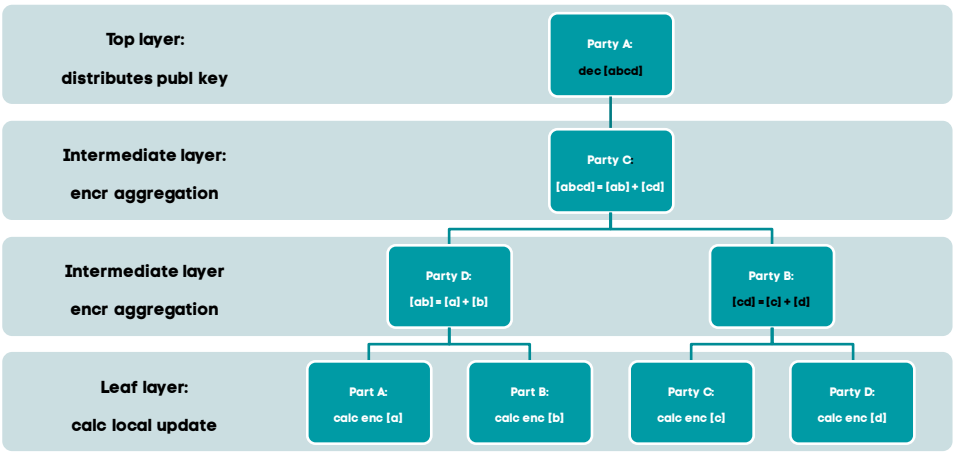

Fig. 1. Binary tree structure where all aggregation is done under Fully Homomorphic Encryption (FHE) to protect model updates of individual parties. Only the global update can be accessed in plaintext. All computations is done in Secure Processing Environments (SPEs), where intermediate results are erased immediately after used. This framework is information symmetric and supports purpose limitations, data- and storage minimization.

that are based on neural networks with many intermediate layers between the input layer and the output (prediction) layer. Deep neural networks have been demonstrated to have great capacity in identifying complex patterns. Model parameters (or weights) are variables that the machine learning algorithm adjusts to optimize its performance. For neural networks, this is often done by gradient descent (or related methods) which iteratively adjusts the parameters of a model to minimize the loss function.

Overfitting occurs when a model becomes too complex and fits the training data too closely, resulting in poor performance on unseen data. Techniques such as regularization, cross-validation, and early stopping can be used to avoid overfitting. Such settings are denoted hyperparameters, which in addition to controlling the behavior of training typically also include parameterizing the machine learning architecture, e.g., the size and depth of a neural network.

Federated learning: Federated learning is a type of machine learning that allows multiple parties to train a shared machine learning model, without directly exchanging their respective data. This makes it suitable for use in scenarios where data is distributed among different parties, and where privacy and security concerns prevent the sharing of data. The case fits well for health- and medical data [6]. For technical reviews of the current and future applications of federated learning for biomedical data, see for example [7]–[9]. Recent experimentation demonstrates that just keeping the data locally is insufficient with regard to the security of the data. Machine learning models are prone to several privacy attacks which could expose sensitive data: 1) An attacker can use the gradient information of the deep learning model to get the sensitive data. 2) Even the trained local model parameters expose information that can be used by an attacker to make an inference about the federated learning participant. The next section goes into some more detail about these types of privacy attacks.

B. Attacks on privacy

1) Membership inference attack: (MI) A membership inference attack aims to determine whether a specific data point has been used during the training of a machine learning model. This method can potentially expose sensitive information about individuals, i.e. whether a person with certain characteristics and a particular medical condition has been included in the model's training data. The attacker can either have black-box access [10], where they only have query access to the model, or white-box access [11], where they have full access to the model's parameters and architecture. Shokri et al. [10] proposed one of the first attacks, which considers an attacker who can query the target model in a black-box way to obtain confidence scores for the queried input. Among the multitude of attack procedures that were proposed later on, we mention [11] that is computationally simpler, but requires that the attacker can calculate the training loss of a candidate data point threshold and compare it with a threshold (the average training loss). A naive baseline procedure was proposed by [12], which predicts a sample as a member if it is correctly labeled by the target model and predicts it as a non-member if misclassified. In a recent experimental comparison [13], the naive model demonstrates similar performance as the more involved MI attack procedures. The two approaches both have a high false positive rate. Indeed, MI attack accuracy is reported to be highly correlated to the model's overfitting or generalization gap [19, 20, 22], and furthermore troubled by high false positive [13]. The generalization gap refers to the difference between the test set and training set performance. As low as possible is generally desired as it reflects the extent to which a model is overfitted. As overfitted models have limited practical use, it is questionable how well reported MI attack success stories can be generalized to well-trained models [14]. Despite the limitations of current MI attack strategies, it is important to study and learn from them as superior attacks might appear in the future.

2) Model inversion attack: The aim of a Model inversion attack is to learn hidden sensitive attributes of a test input given knowledge about the non-sensitive attributes. This attack is also called an attribute inference attack and is carried out as a search for the value of the sensitive attributes that maximizes the posterior probability given the non-sensitive attributes, model access, and prior knowledge about the distribution of attributes [15]. This attack exploits the correlation between the sensitive attribute

TABLE I
PRIVACY ATTACKS CONSIDERED FOR THIS WORK.

Membership inference attack Model inversion attack (attribute inference attack)	• Infer whether a specific data point has been used during the training of a machine learning model. • Infer hidden sensitive attributes of a training input given knownledge about the non-sensitive attributes. • Both attacks use similar procedure and input data (which is why they are often treated together). • A party in federated learning could use its own data and a global model to learn about other parties' data. • Reported successful attacks may rely on model overfitting. How relevant is this for more robust models? • Although perhaps not practically possible today, superior attack procedures may appear in the future.
Gradient inversion attack	• Practical to reconstruct data points from gradients averaged over several iterations or batches. • Successful attacks recovered single data points from batches of up to a hundred images or texts.

and the model output, which is encoded in the machine-learning model. Many of the proposed attack procedures are modified variations of membership inference attacks, for example, [11], why it often makes sense to discuss both these attacks together. Also related, are the memorization attack, which exploits the ability of high-capacity models to memorize certain sensitive patterns in the training data [16]; and the property inference attack, in which the attacker tries to infer whether the training data set has a specific property. Although these attacks are related to attribute inference, it is rather the overall statistical patterns of the training data that are exposed. As the topic of our discussion concerns the privacy of the individual, it is sufficient to consider attribute inference.

3) Gradient inversion attack: The gradient used to improve the model contains information about the batch of data points that were used to calculate it. Early work that recovered data from gradient information was limited to shallow networks of less relevance. Later, it was shown to be [17] possible to reconstruct up to 8 images from their batch averaged gradients also for slightly deeper neural networks. More recently, [18] explored settings encountered in practice when training deep neural networks and showed that even averaging gradients over several iterations, or several images, does not protect the privacy of an individual data point in federated learning applications. Indeed, by exploiting a magnitude-invariant loss function, it is possible to faithfully reconstruct images at high resolution from their parameter gradients for realistic deep architectures like ResNet. The reconstruction is possible even when averaging gradients over multiple epochs, using local mini-batches, or even for a local gradient averaging of up to 100 images with deep networks, appearing to be as vulnerable as shallow networks. Attacks against federated averaging of parameters (instead of gradients) have also been devised [19].

C. Privacy enhancing technologies

1) Homomorphic encryption: Fully homomorphic encryption (FHE) allows mathematical operations to be performed directly on encrypted data, without first decrypting the data, and without access to a secret key. FHE is distinguished from conventional uses of cryptography, where data is encrypted only while it is sent between parties (in motion), and during storage on a file system (at rest) but is decrypted for calculation and processing. This last step of decryption introduces a vulnerability to conventional cryptography, in that data can be read in hardware or software layers, and that a secret key must be available on the server that performs the calculations. FHE offers a solution that guarantees that even a curious computing party can not see the data. It enables privacy-preserving processing and analysis of data, for example in a cloud-based AI service (Figure 2), where the original data as well as all intermediate and final results are indistinguishable from random noise to the computing cloud.

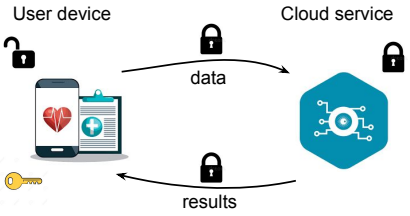

Fig. 2. A two-party solution that uses homomorphic encryption to protect data sent from a user device to be processed in the cloud. (Image from [20])

Decryption is possible only for the private key holder and considered unbreakable under very strong cryptographic guarantees for most recent FHE schemes, e.g. even against hypothetical quantum computer-based attacks [21]. We use the word *plain text* for unencrypted data and *cipher text* for encrypted data, as is conventional.

Fully Homomorphic Encryption differs from early schemes referred to as Partially Homomorphic Encryption schemes in that FHE can support both addition and multiplication. In this work, the difference in terminology is not important as we only consider schemes with full arithmetic support, and will refer to them either as (Fully) Homomorphic Encryption or simply abbreviated as FHE. Noise is added in the construction of the encrypted representation. For every operation, this noise grows such that the result of large computations can be meaningless after decryption unless noise-mitigating measures like bootstrapping are used. In a leveled

approach one carefully manages the cryptographic *noise-budget*, by which we mean the number of arithmetic operations one can carry out before bootstrapping becomes necessary. Practically this means that only a limited number of multiplications and additions are allowed. This is controlled by the parameters selected for encryption, which we will refer to loosely as "key size".

The CKKS scheme [22] named after its developers Cheon, Kim, Kim, and Song, encrypts complex numbers and can perform fixed-point arithmetics. Its security is based on the Ring Learning With Errors (RLWE) problem. While many other schemes perform exact arithmetics on encrypted integers, CKKS has become popular for applications that only require approximate calculation. The TFHE scheme developed by Chillotti and collaborators [23] supports fast bootstrapping, arithmetic operations, and univariate function evaluation thanks to its implementation of a type of look-up mechanism. An important difference between the software libraries used in this work is that the CKKS implementation [24] supports SIMD operations, while the TFHE library [25] does not. This means that the former can add (or multiply) vectors up to a certain size at constant cost, while for the latter the cost of additions of vectors is linear in the length of the vector.

Proposition for using FHE to enhance security in Federated Learning applications have recently been put forward [26]–[29]. This provides efficient defense against the above-mentioned attacks by only allowing the exchange of encrypted information between the participants of the federation such that it can be aggregated (i.e. averaged) under homomorphic encryption. However, it comes with significant overhead in terms of computation time and data transfer.

2) Differential privacy: Differential privacy is a framework for privacy-preserving data analysis, where a randomized algorithm $\mathcal{A}$ is considered (ϵ, δ)-differentially private [30] if for any neighboring datasets D_1 and D_2 that differ, in at most, one record and for any set of outputs $S \subseteq Range(\mathcal{A})$,

$$P[\mathcal{A}(D_1) \in S] \leq e^{\epsilon} P[\mathcal{A}(D_2) \in S] + \delta,$$

where δ represents the maximum allowable probability that privacy is violated. In other words, the (ϵ,δ)-differential privacy guarantee provides a limit on the overall probability of privacy violation.

The noise that is added to a differentially private algorithm's output is calibrated based on the sensitivity of the function being computed, which is defined as the maximum distance in some norm, $||\cdot||$, between the outputs of neighboring data sets, $\Delta f = \max_{D,D'} ||\mathcal{A}(D) - \mathcal{A}(D')||$.

The Rényi mechanism [31] is a variant of differential privacy that can be useful for machine learning applications as it allows for fine-grained control over the level of privacy while maintaining the accuracy of the output also over compounded applications. It can be used to perturb training data or model updates, thereby providing a privacy-preserving mechanism for training machine learning models on sensitive data also in a sequential, iterated application like gradient descent.

It is common to add noise as a perturbation to the gradients in differential privacy applications for machine learning. To ensure that the added noise is proportional to the sensitivity of the model, the gradients are often clipped before the noise is applied. This involves constraining the magnitude of the gradients to mitigate the effect of high sensitivity. By controlling the amount of perturbation and clipping, one can achieve a trade-off between privacy and model accuracy. Questions have been raised about real-world applications using very high epsilon to achieve utility over composition [14], although recent work points to more favorable trade-offs, e.g., for Stochastic Gradient Descent with noise [32].

3) Secure aggregation: Secure aggregation is a multi-party computation technique enabling non-trusting parties with sensitive data to privately compute an aggregate without depending on a trusted third party. This process typically involves the following steps: i) clients agree on pairwise private seeds, ii) each client generates a private seed, iii) clients randomly mask their model parameters using the seeds and communicate the masked model to the server, and iv) clients distribute shares of the seeds to other clients using a secret sharing scheme [33], [34]. The secret sharing is based on a (t, n) secret sharing scheme and offers resilience against dropouts and stragglers, i.e., clients not responding to the server, by allowing the random seeds of each client to be recovered from the collected shares of t out of the n clients. This property is leveraged during aggregation where the server requests shares from the available clients to reconstruct the sum of the secret masks so they can be canceled out.

Since secure aggregation occurs over a finite field, clients must convert their model parameters accordingly. The size of the finite field can impact the model utility of secure aggregation; larger field sizes preserve model utility but increase communication overhead whereas a smaller field size may result in loss of information. Secure aggregation generally incurs extra communication compared to differential privacy and homomorphic encryption.

Recently, researchers have combined secure aggregation with differential privacy to mitigate the negative impact on model utility caused by differential privacy [35]. The core concept is to protect the aggregate of local models rather than individual local models, resulting in the addition of less noise and, ultimately, a lesser effect on model utility.

D. Regulatory environment

The purpose of GDPR and other privacy legislation is to protect individuals' personal data and privacy rights by establishing clear principles for the collection, use, and processing of personal data. Important principles include the lawful, fair, and transparent processing of personal data, purpose limitation, data minimization, accuracy, storage limitation, integrity and confidentiality, accountability, and respect for individual rights. In this article, we will particularly consider:

Purpose limitation: Personal data must be collected for specified, explicit, and legitimate purposes and not processed in a manner that is incompatible with those purposes.

Data minimization: Personal data must be adequate, relevant, and limited to what is necessary for the purposes for which it is processed.

Storage limitation: Personal data must not be kept for longer than necessary for the purposes for which it is processed.

This work will now discuss how FHE and other PETs can support the above general principles, and be relevant mitigations to consider also in a specific use-case.

Changing environment: The Swedish government has launched official investigations [36], [37] in order to achieve a national data strategy aimed at increasing the access and beneficial utilization of data. Such purposes include improved health applications supported by artificial intelligence. During 2021-22, The Swedish Authority for Privacy Protection (IMY) was commissioned by the government to provide support and guidance to the innovation system on data protection matters [38] Related to this mission, IMY organized activities where experts and participants from industry and public sector could interact. These included research hearings on PETs, workshops, and seminars.

A recent report [39] commissioned by the Swedish eHealth Agency explores the benefits of a national data space for medical AI, particularly in image diagnostics and mammography. It examines the concept of a Secure Processing Environment (SPE), where data can be isolated and encapsulated to prevent unauthorized access and protect sensitive information while still allowing researchers and healthcare professionals to use the data for research and analysis purposes. The report calls for deeper investigations of how federated learning in a distributed ecosystem of SPEs can facilitate the safe sharing of resources and data, and increase opportunities for research and innovation while meeting important integrity- and legal requirements.

The Swedish innovation agency Vinnova, in a recent report commissioned by the government [40], discussed various aspects of secure data sharing, including the need for increased dissemination and utilization of conventional privacy protection techniques. It also highlights the importance of conducting research on cutting-edge technologies, especially mentioning federated learning and homomorphic encryption. Also, AI Sweden, the national center for applied artificial intelligence, views decentralized AI as one of the critical technologies for future AI development [41] across several business and industrial sectors. Although new advanced technologies for privacy protection hold much promise regarding, on the one hand, legal and security requirements, and on the other hand, exploiting the potential of data sharing and utilization in health and medicine in Sweden, legal uncertainty still remains.

Regulatory sandboxes: Regulatory sandboxes aim to bridge the gap between the rapid pace of technological development and the slower pace of regulatory and policy development. They can assist in identifying and developing new ways of working in the public sector that could enable more agile and effective regulatory and policy responses to emerging challenges. By engaging with innovators and working together to identify legal ambiguity and challenges, governing bodies that participate can promote a more effective and efficient regulation that supports innovation while still protecting the public interest.

Regulatory sandboxes have already been put in place in the UK, Norway, and France with guidance that targets the application of GDPR. The goal is to increase judicial predictability, reduce time and risk for a product or service to reach the market, and facilitate startup and small business growth by doing so. In the EU [42], regulatory sandboxes have been highlighted as a way to promote innovation and growth for companies, and the draft AI regulation being negotiated currently includes proposals for regulatory sandboxes to promote and facilitate the application of AI. Regulatory sandboxes allow for exploratory, dialogue-based guidance to be given to selected innovation projects in exchange for the work being summarized in a public report that enables learning for others. The approach helps to develop practical examples in areas where both technology and law are complex, relatively new, and untested, while also increasing regulatory authorities' understanding of new technology and how it can be applied.

Sandbox: Decentralized AI in Healthcare: IMY participated in a pilot project on regulatory sandboxing in 2022 and summarized its conclusions in a public report [43]. The project, titled "Decentralized AI in Healthcare - Federated Machine Learning between Two Healthcare Providers" focused on evaluating the potential of jointly training and exchanging machine learning models between two healthcare providers, Region Halland and Sahlgrenska University Hospital, in order to predict heart failure patient readmissions within 30 days of their last hospital stay. The project was facilitated by AI Sweden, the national center for Applied AI.

The purpose of the project was to explore the potential of regulatory sandboxing as an approach to address complex societal challenges and to help regulators and policymakers better understand and analyze new technologies that fall within their regulatory frameworks. Specifically, the project aimed to provide in-depth guidance on how data protection regulations should be interpreted and applied to a specific innovation initiative involving advanced technologies like AI and federated machine learning.

The following paragraphs summarize the parts of the report which are important for our discussion in this article, starting with its three focus questions:

Question 1: Is there a legal basis for local processing of personal data, i.e., when healthcare providers train the machine learning model locally only on their own patient data? IMY's assessment is that there is a legal basis for local processing of personal data. The key factor is that IMY believes there is support for a dynamic and technology-neutral interpretation of the purpose provisions in the Patient Data Act and the Health and Medical Services Act, which means that what falls within these provisions can change over time, with regard to technological development.

Question 2: Does personal data disclosure occur between healthcare providers in the federated machine learning in this case? IMY's assessment is that Region Halland and Sahlgrenska University Hospital are at risk of disclosing personal data to each other in the current case when the knowledge gained from local training is combined into a joint machine learning model. Either party could, if it gathers the necessary expertise and purposeful intent, launch two types of privacy-harming attacks, namely, Membership Inference Attack and Model Inversion Attack to infer information about persons in the other party's data set.

Question 3: Is there a legal basis for disclosure of personal data between healthcare providers? IMY has not made any assessment of whether any confidentiality-breaking provision could be applicable in the current case. However, if Region Halland and Sahlgrenska University Hospital, both being authorities, were to request patient data from each other with the support of the Public Access to Information and Secrecy Act, such disclosure could possibly be allowed provided that the data is not confidential. However, patient data within healthcare is generally confidential. The legal basis for personal data disclosure between healthcare providers under certain circumstances may

be in place but requires a case-by-case assessment.

IMY limited their investigation to the use case at hand: federated learning between two public health care providers. As a precaution in the project, the two healthcare providers only used data that did not contain personal information.

TABLE II
INFORMATION ASSUMPTIONS IN FRAMEWORK

Trivially known by all parties • Model architecture • Training hyperparameters such as learning rate, regularization (gradient clipping, etc) • The aggregation topology (tree)
Each party knows at the end • Content of their own data set (1) • The final model parameters (2)
Each party at each iteration • Their own (local) model update (3a) • The global model update (3b)
The central party • Knows and holds the secret key • Distributes the public keys

III. METHOD

A. Problem statement

Here we restate the objective from the Introduction: How can we design privacy-enhancing measures for a federated learning effort such that privacy is best preserved without unduly compromising the utility of the trained model? We also limited the exploration in this work to two privacy problems:

- P1: the parameters of an AI model can potentially leak personal data, and
- P2: in the federated learning process parties can potentially leak information about their data set through the (iterated) exchange of model updates.

The IMY report primarily discussed (P1) in how it opens up vulnerability to membership inference attacks, and model inversion attacks. Here we also want to highlight that (P2) should be considered carefully as it has been demonstrated that inference of private data from model gradients is practical for certain machine learning models and settings [18].

B. Target Framework

The overarching intention of the federated learning framework discussed below is to promote data- and storage minimization, purpose limitations, and symmetric distribution of information, such that every party is trusted only with the data it needs to perform a task according to the original intentions while avoiding trust asymmetries where some party has privileged access to the information of others.

Preparaion stage: One of the parties in the federation is selected to create a secret key which is used to produce the public keys that it distributes to the other parties.

1) Secure processing environments: First and foremost we will assume that each party carries out all computations in a Secure Processing Environment under strict access control. They are furthermore obliged by contractual agreements to follow the machine learning protocol and take appropriate storage minimization measures, like deleting intermediate model updates immediately after they are used.

2) Homomorphically encrypted: Homomorphic encryption is used throughout the federated learning process to encrypt individual parties' model updates with a public key that they have received before the model training exercise. This contributes towards the goals of data minimization and purpose limitation.

3) Aggregation over binary-tree: Encrypted model updates are aggregated between parties in a binary tree structure of Figure 1. It is strictly not necessary to use a binary tree, as long as the top party receives only the encrypted global aggregate, which it decrypts and distributes to all other parties. This removes the information advantage of the party that holds the secret key and decrypts the global model update. Every party now only has access to its own model update and the aggregated global model.

4) Differential privacy: Differential privacy has the potential to protect against both problems P1 and P2, but one carefully has to consider for each particular use case if it has adverse implications for utility. Even if one does not promise full differential privacy, the addition of noise at a higher value of ϵ (together with other regularization) can help avoid overfitting, which also combats privacy attacks.

On completion: At the end of the training effort, all hardware that has touched the data is thoroughly erased (or even destroyed). The key holding party must similarly permanently delete the decryption key.

Information symmetry: Table II summarizes the information that each party knows throughout the exercise.

C. Models and Materials

Numerical experiments were carried out to investigate the performance of differential privacy and federated learning that uses homomorphic encryption during parameter aggregation. For these experiments we used two different models chosen such that both a very simple as well as a moderately complex architecture were examined, that is a logistic regression model (LogReg) and a deep learning model for image analysis (ResNet-18) whose features are both summarized in Table IV.

1) Logistic Regression: A logistic regression model was trained to estimate the risk of future coronary heart disease (CHD) based on a patient's information such as demographic, behavioral, and medical factors. The dataset is publicly available on the Kaggle website [44], and it is from an ongoing cardiovascular study on residents of the town of Framingham, Massachusetts. It includes over 4,000 records and 15 attributes, from which a balanced subset of 1,000 records and 8 attributes was selected (such that positive and negative labels were equally frequent). A test set of 200 records was set aside, while the remaining 800 records were used for training the logistic regression model.

2) Deep Neural Network: ResNet-18 is a large image classification based on deep neural networks, which is described later in this section. It was trained to classify images from

TABLE III
DP ANALYSIS STATS.

	inf	0.1	0.3	1.0	3.0	10.0	30.0
LogReg	68.9±0.9	67.3±2.2	68.1±1.6	67.9±1.5	68.5±1.6	68.6±1.5	68.2±1.5
ResNet	64.7±2.1	19.2±1.1	29.7±1.1	35.6±0.4	39.6±0.0	41.5±0.2	45.1±0.1

TABLE IV
SUMMARY OF MODELS

	LogReg	ResNet
total params	10	11511784
train params	10	3591
data set name	FraminghamCHD	DermaMNIST
data size used	1000	10015

TABLE V
MEAN TEST SET ACCURACY FOR LOGREG (WITH SAMPLE STD).

	2	4	8	16
tfhe 256	68.6±1.3	68.1±1.6	64.1±4.5	58.3±6.1
tfhe 512	68.8±1.1	68.8±0.9	68.9±0.8	69.4±0.7
plain text	68.8±1.0	68.7±1.0	69.0±0.9	69.3±0.8
ckks 4096	68.7±0.9	68.8±0.8	69.1±0.9	69.2±0.7
ckks 8192	68.8±1.0	68.9±1.0	69.1±0.9	69.2±0.8

the DermaMNIST/HAM10000 [45] collection of multi-source dermatoscopic images of common pigmented skin lesions. The dataset consists of 10,015 dermatoscopic images categorized as 7 different diseases and was downloaded using the MedMNIST software library [46], [47].

Training Procedure: About 20% of the data for each model was set aside as a test set. The remaining 80% of the data was used for training. The models were trained for a total of 5 epochs.

Differential Privacy: Each model was trained under the Rényi mechanism for compounded (ϵ,δ)-differential as implemented in the Opacus [48] library for PyTorch. Standard settings were used for δ and gradient clipping, while ϵ was varied over a fixed range from $\epsilon=0.1$ (relative strong privacy) to $\epsilon=30$ (weak privacy).

Federated Learning: For each model, a federated learning set-up was simulated that at the end of each epoch, encrypted parameters were aggregated across the nodes organized in a binary tree-like topology (Figure 1), with the number of nodes taken as $n \in \{2,4,8,16\}$. The training data was split evenly between the nodes that were part of the federated exercise. Each model was trained using the two different homomorphic encryption schemes, TFHE and CKKS, each with two different parameter settings. For comparison, each model was also trained with plain text aggregation for each node configuration. The training was repeated 100 times to gather statistics about the variation in performance.

IV. RESULTS

A. Differential Privacy Experiments

Table III reports the mean accuracy with 95% confidence bands for each examined machine learning model. Each column represents a target ϵ, where the first column reports the case of no differential privacy, i.e., when no noise was added. For the LogReg model, we see that the estimated mean accuracy falls within the confidence bands of the non-private model for all but the lowest ϵ values. There doesn't seem to be a significant adverse effect on utility from adding differential privacy for this case. The results are very different for the deep learning model. The estimated mean accuracies for all the differentially private ResNet models are far below the lower confidence bound for the non-private model.

B. Federated Learning Experiments

1) Logistic regression model.: Table V displays the mean accuracy obtained for the logistic regression model, together with the observed standard deviation across the 100 repetitions of the experiment. The accuracy appears relatively unaffected by the encrypted aggregation, except for the smaller key size for the TFHE scheme for the size $n=8$ and $n=16$ node configurations where mean accuracy is more than 2 std worse than for the corresponding plain text configuration. For all other model configurations, the results using homomorphically encrypted aggregation are not significantly different from the plain text case.

Execution times were also collected throughout the numerical experiments. For the LogReg model, the measured average time in milliseconds for the central cryptographic operations is displayed in Table VI. For both schemas, encryption of the model parameters ("enc time") dominates both the time it takes to carry out the additions ("add time") and the time it takes to decrypt the aggregated results ("dec time"). Here TFHE appears to be overall faster, although we note that CKKS outperforms for the addition.

TABLE VI
TIMED OPERATIONS FOR LOGREG IN MS.

context	enc time	add time	dec time
tfhe 256	0.334	0.088	0.015
tfhe 512	0.623	0.162	0.028
ckks 4096	4.03	0.059	0.962
ckks 8192	10.071	0.149	2.916

2) ResNet-18 model.: ResNet-18 is a convolutional neural network that is 18 layers deep [49] and has more than 11 million parameters. We used a version of the network that was pre-trained on more than a million images from the ImageNet database [50] on the task to classify images into 1000 object categories, such as keyboard, mouse, pencil, and many animals. The network has thus learned a rich feature representation for a wide range of images, which is can be leveraged for the medical image classification task. When data is scarce, one can sometimes take a model trained on data from a related task and then fine-tune the model by training it on the target task, which is an example of transfer learning (see

TABLE VII
MEAN TEST SET ACCURACY FOR RESNET-18 (WITH SAMPLE STD).

scheme	2	4	8	16
tfhe 256	68.1±0.7	67.3±0.8	66.3±0.9	64.7±1.5
tfhe 512	68.4±0.6	68.0±0.7	67.5±0.5	67.2±0.5
ckks 4096	68.3±0.6	67.9±0.5	67.6±0.5	67.1±0.5
ckks 8192	68.4±0.6	68.1±0.5	67.6±0.5	67.2±0.5
plain text	68.5±0.6	67.9±0.6	67.5±0.5	67.2±0.5

TABLE IX
FILE SIZE (KB) FOR DIFFERENT PARAMETER SIZES.

context	1	10	100	3591
tfhe 256	5.458	53.629	535.323	19223.441
tfhe 512	10.682	105.88	1057.892	37988.886
ckks 4096	80.897	80.908	80.904	161.809
ckks 8192	334.32	334.298	334.314	334.314
plain text	0.025	0.255	2.549	91.55

for example a recent review [51]) Thus we can obtain acceptable performance after only 5 epochs of training on a relatively small data set of 8007 images of skin changes. In fact, we only let the 3591 parameters of the top layer be trainable, and keep all other layers at their pre-trained values.

Similar results as for the logistic regression case are observed also for the accuracy of the ResNet18 model trained on the DermaMNIST data. Table VII shows an impairment of the accuracy of at least two standard deviations for the configurations with the smaller key size of the TFHE scheme. For the other configurations using homomorphic encryption aggregation the difference compared to the plain test case is not significant.

TABLE VIII
TIMED OPERATIONS FOR RESNET-18 IN MS.

scheme	enc time	add time	dec time
tfhe 256	46.1	128.4	1.9
tfhe 512	85.1	309.5	4.1
ckks 4096	8.0	0.6	2.0
ckks 8192	17.2	1.3	3.0

Table VIII displays the execution times for the operations that support the homomorphic encryption aggregation. Here, performance is reversed such that CKKS outperforms TFHE, since the former supports SIMD (Single Instruction Multiple Data) execution, meaning that the summation of parameters from two different nodes can be "vectorized", i.e. carried out in parallel over the entries in the same position as the encrypted representation can hold vectors. Thus we can benefit from SIMD when summing the 3591 parameters of the ResNet-18 top layer from two nodes. At the time that we carried out the experiments the TFHE software implementation [25] that we used did not support SIMD and thus had to encrypt each individual entry of a vector separately. It was supported by the CKKS implementation [24] that we used.

Data size overhead: The encryption inevitably results in storage and communication overhead since the encrypted representations are larger than the plain text. This effect is very noticeable for the two encryption schemes considered in this work, which is evident in Table IX that lists the file storage size in kilobytes (KB). Column headers identify the number of stored values, i.e. the length of a vector of real numbers. Columns with labels 10 and 3591 lists the size of the representations for the LogReg and ResNET-18 models, respectively. We note that data size requirements are lower for THFE for the LogRes model with only 10 parameters, while the requirements for ResNet are (much) smaller for CKKS than for TFHE (last column).

CKKS natively stores vectors and have more compact representations, which are in fact invariable to the size of the plain text vector for lengths up to half of the key size (poly mod); therefore CKKS 8192 can store the entire vector of trainable parameters for ResNet-18 in a single file since it has length 3591 which is smaller than $8192/2 = 4096$.

For the smaller LogReg model, the data size is increased by about three orders of magnitude compared to the plain text. For the larger model (ResNet-18) CKKS outperforms thanks to its compact representation and only shows a modest overhead compared to the plain text.

V. DISCUSSION

The healthcare providers that participated in IMY's regulatory test operation were data controllers for local patient data processing. The setup in the IMY study required a central party that was given plain text access to all the model updates at each iteration. This was considered a potential transfer of personal data, however, IMY did not assess whether there is a legal basis for a healthcare provider to process personal data originating from another healthcare provider.

Other examples of issues that were not considered in the pilot project, including how the right to information of the registered individuals should be met and the question of the data controller's requirement not to handle more personal data than necessary (the principle of data minimization).

Information symmetry: In the federated learning framework proposed in this article based on encrypted aggregation in a tree-like structure, all parties have symmetric information access. There is thus no party with privileged access to the plaintext model updates of all other parties. Each party knows its own local data and model update, as well as the corresponding global information (recall Table II).

A curious participant can additionally subtract their own model update from the global update, to learn the aggregate model update of all other parties (except themselves). Because of the encrypted aggregation, they do not directly access the model update from any individual party. In a set-up with more than two parties, the information from a single party is now blended with that of many other parties.

With a larger aggregate batch size (in the hundreds of data points), gradient inversion attacks like [18] should be difficult to launch successfully, especially if they target a single party's data that is now diluted in the aggregate. The addition of noise in the training process should make it harder, even if full differential privacy may not always be appropriate because of the utility-privacy trade-off.

Data minimization: The minimal information that is needed by each party for improving the model (in addition to their own

data) is the global model update. This is the only (non-trivial) information that is shared in plain text in the proposed framework. If one further requires these to be deleted after used to update the model, one additionally meets *storage minimization* criteria.

Purpose limitations: During model training, the encrypted model updates that are sent up in the binary tree have to be added, but this is also the only meaningful use (for a non-malicious participant as in our trusted-but-curious set-up). The parties that participate in the exercise, have agreed not to make unintended use of the information they are trusted with; however, data remaining on the system could be used by a future actor with other intentions. Therefore model updates should be deleted after use. The final parameters of the model are, however necessary to maintain for the deployment of the model.

Deployment: A party must know the global model parameters for deploying the resulting model on its own system. Model parameters are thus to be considered part of the minimal set in a self-deployment scenario. It is, however not necessary for every party to keep the model parameters if the model is run by a single member of the federation. This opens new privacy concerns when data is sent for inference to the centrally hosted model, which can be mitigated by the use of homomorphic encryption for the transferred data. A remaining concern in such a setup may be that the model hosting party must be trusted with plain-text access to the model parameters.

In an effort of extreme data minimization, each party could encrypt model parameters with their own secret key and outsource the running of the model to a third party (that does not have access to the decryption key). They then permanently erase all traces of the model parameters (and model updates) in their own systems and only use the hosted encrypted models, in each request sending it data encrypted with their own key. They must maintain the secret key in order to decrypt the returning results. Albeit computationally expensive, such a solution mitigate membership inference attacks and model inversion attacks by preventing an attacker from knowing the model parameters (white-box), reducing opportunities to a much harder black-box access-only scenario. However, because of the high computational cost, one should first consider plain text deployment in a secure processing environment where data is protected with conventional encryption while being transferred between client and server in the organization.

Scaling: It may be preferable to use a scheme that supports SIMD operations, as the ability to encrypt vectors and carry out parallel element-wise addition greatly improves performance for larger models. This was illustrated in this work by the comparison of training the smaller logistic regression model and the larger ResNet model. The performance versus accuracy trade-off was more favorable for CKKS as this scheme supports SIMD operations and compact representation of an encrypted vector, which reduces both the computational effort and the data file size. For the federated learning under FHE experiment, only the top layer (of 3591) parameters of ResNet-18 were trained, which could fit within two ciphertexts for CKKS with the smaller key size. If we were to train all layers in the network with 11511784 parameters, it would instead require 5621 cipher texts. Assuming linear scaling, the computation time for adding the numbers would grow to 3.4 s (although the operations could be parallelized over the ciphertexts for faster execution). Similarly, the file size would grow to almost 460 MB. For TFHE, this would require one cipher text per parameter, with clearly worse scaling of time and memory cost by factors of hundreds compared to the CKKS estimates.

Total overhead: If we exclude the loading of keys and other operations that do not cause repeated overhead and only consider the overhead for using homomorphic encryption that is part of each federated learning iteration, we have

1) τ_{enc}: encryption of parameter vectors
2) τ_{add}: addition of encrypted vectors
3) τ_{com}: transfer of encrypted vectors
4) τ_{dec}: decryption of aggregated vector

Of these, we count (1) only once as it is carried out in parallel across all parts of the federation. The overheads from (2) and (3) are counted once per level in the aggregation tree, as they happen simultaneously for all participants at that level, and hence have an impact logarithmic in the size n of the federation. Finally, (4) is done only by the aggregating party and also only happens once. The total overhead then be estimated as

$$\tau\mathrm{tot} = \tau_{\mathrm{enc}} + \tau_{\mathrm{add}}\log_2 n + \tau_{\mathrm{com}}\log_2 n + \tau_{\mathrm{dec}},$$

where we have assumed that the data processing and transport is synchronized between the parties to avoid lag.

Legal uncertainty: Homomorphic encryption, differential privacy, and federated learning are not well covered by existing privacy laws and regulations, and their use can raise questions about compliance and liability. On the other hand, it can help manage and mitigate legal and regulatory risks by providing organizations with more robust and verifiable mechanisms for complying with privacy laws. For example, an organization can provide evidence of its efforts to protect personal data. Furthermore, the combined use of the technologies may enable organizations to collaborate on new research and development projects that would otherwise not be feasible. Such efforts could improve AI-based treatment methods that lead to better health outcomes as well as commercial opportunities without compromising the privacy of the individuals who ultimately contributed the data.

VI. Conclusions

Numerical experiments where two standard text-book models were trained both under differential privacy and in a federated learning set-up that uses homomorphic encryption for model parameter aggregation. The experiments confirmed that differential privacy can have a significant adverse impact on utility for some training scenarios. For the use of FHE, it was concluded that the approach was feasible not only for the trivial regression model but also for the more advanced deep learning model. Of the two FHE schemes tested, the performance versus accuracy trade-off was more favorable for CKKS as this scheme supports SIMD operations and compact representation of an encrypted vector, which reduces both the computational effort and the data file size.

Future Work: We would like to extend the numerical experiments to also compare the framework based on FHE with alternatives that use secure aggregation in combination with differential privacy. We will continue to explore implementation together with Swedish healthcare providers to gain a roadmap to practical PET application.

ACKNOWLEDGMENT

We thank Dr. Magnus Clarin, Head of Research and Education, Region Halland, and Dr. Magnus Kjellberg, Head of AI Competence Center at Sahlgrenska University Hospital, Västra Götalandsregionen, for their contribution to the work that is described in this manuscript.

REFERENCES

[1] Schork NJ. Artificial intelligence and personalized medicine. In: Precision medicine in Cancer therapy. Springer; 2019. p. 265-83.
[2] Fitzpatrick F, Doherty A, Lacey G. Using artificial intelligence in infection prevention. Current treatment options in infectious diseases. 2020;12(2):135-44.
[3] Olatunji IE, et al. A Review of Anonymization for Healthcare Data. arXiv preprint arXiv:210406523. 2021.
[4] Torkzadehmahani, et al. Privacy-preserving AI techniques in biomedicine. Methods of Inf in Med. 2022.
[5] Gentry C. Computing arbitrary functions of encrypted data. Comm of the ACM. 2010 mar.
[6] Rieke N, et al. The future of digital health with federated learning. npj Digital Medicine. 2020 sep;3(1).
[7] Xu J, Glicksberg BS, Su C, Walker P, Bian J, Wang F. Federated Learning for Healthcare Informatics. Journal of Healthcare Informatics Research. 2020 nov;5(1):1-19.
[8] Antunes RS, da Costa CA, Küderle A, Yari IA, Eskofier B. Federated Learning for Healthcare: Systematic Review and Architecture Proposal. ACM Trans on Intelligent Systems and Tech. 2022;13(4).
[9] Kairouz P, et al. Advances and Open Problems in Federated Learning. Now Publishers; 2021.
[10] Shokri R, Stronati M, Song C, Shmatikov V. Membership Inference Attacks Against Machine Learning Models. In: 2017 IEEE Symposium on Security and Privacy (SP). IEEE; 2017. .
[11] Yeom S, Giacomelli I, Fredrikson M, Jha S. Privacy Risk in Machine Learning: Analyzing the Connection to Overfitting. In: 2018 IEEE 31st Computer Security Foundations Symp (CSF). IEEE; 2018. .
[12] Leino K, Fredrikson M. Stolen Memories: Leveraging Model Memorization for Calibrated White-Box Membership Inference. In: Proc of the 29th USENIX Conf on Security Symp. SEC'20. USA; 2020. .
[13] Rezaei S, Liu X. On the Difficulty of Membership Inference Attacks. In: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE; 2021. .
[14] Jayaraman B, Evans D. Evaluating Differentially Private Machine Learning in Practice. In: Proceedings of the 28th USENIX Conference on Security Symposium. SEC'19. USA; 2019. p. 1895–1912.
[15] Fredrikson M, Eric Lantz SJ, Lin S, Page D, Ristenpart T. Privacy in Pharmacogenetics: An End-to-End Case Study of Personalized Warfarin Dosing. In: Proceedings of the USENIX Security Symposium. USENIX; 2014. .
[16] Carlini N, Liu C, Erlingsson U, Kos J, Song D. The Secret Sharer: Evaluating and Testing Unintended Memorization in Neural Networks. In: Proceedings of the 28th USENIX Conference on Security Symposium. SEC'19. USA: USENIX Association; 2019. p. 267–284.
[17] Zhu L, Han S. Deep Leakage from Gradients. In: Lecture Notes in Computer Science. Springer International Publishing; 2020. p. 17-31.
[18] Geiping J, Bauermeister H, Dröge H, Moeller M. Inverting Gradients - How Easy is It to Break Privacy in Federated Learning? In: Proceedings of the 34th International Conference on Neural Information Processing Systems. NIPS'20. Red Hook, NY, USA; 2020. .
[19] Dimitrov DI, Balunovic M, Konstantinov N, Vechev M. Data Leakage in Federated Averaging. Trans on Machine Learning Research. 2022.
[20] Brännvall R, et al. Homomorphic encryption enables private data sharing for digital health: winning entry to the Vinnova innovation competition Vinter 2021-22. In: Proceedings of the Swedish Artificial Intelligence Symposium (SAIS); 2022. .
[21] Augot D, et al. Initial recommendations of long-term secure post-quantum systems; 2015. .
[22] Cheon JH, et al. Homomorphic Encryption for Arithmetic of Approximate Numbers. In: Advances in Cryptology – ASIACRYPT 2017. Springer International Publishing; 2017. p. 409-37.
[23] Chillotti I, et al. TFHE: Fast Fully Homomorphic Encryption Over the Torus. Journal of Cryptology. 2019 apr;33(1):34-91.
[24] TenSeal library implements CKKS for the Python language;. OpenMined. Accessed: 2023-03-30. https://github.com/OpenMined/TenSEAL.
[25] Concrete library implements TFHE for the Rust langauge;. ZAMA. Accessed: 2023-03-30. https://concrete.zama.ai.
[26] Phong, et al. Privacy-Preserving Deep Learning via Additively Homomorphic Encryption. IEEE Trans on Information Forensics and Security. 2018 may.
[27] Ou W, Zeng J, Guo Z, Yan W, Liu D, Fuentes S. A homomorphic-encryption-based vertical federated learning scheme for rick management. Computer Science and Information Systems. 2020;17(3):819-34.
[28] Zhang C, Li S, Xia J, Wang W, Yan F, Liu Y. BatchCrypt: Efficient Homomorphic Encryption for Cross-Silo Federated Learning. In: Proceedings of the 2020 USENIX Conference on Usenix Annual Technical Conference. USENIX ATC'20. USA; 2020. .
[29] Park J, Yu NY, Lim H. Privacy-Preserving Federated Learning Using Homomorphic Encryption With Different Encryption Keys. In: 2022 13th International Conference on Information and Communication Technology Convergence (ICTC). IEEE; 2022. .
[30] Dwork C. Differential Privacy: A Survey of Results. In: Lecture Notes in Computer Science. Springer Berlin Heidelberg; 2008. p. 1-19. Available from: https://doi.org/10.1007/978-3-540-79228-4_1.
[31] Mironov I. Rényi Differential Privacy. In: 2017 IEEE 30th Computer Security Foundations Symposium (CSF). IEEE; 2017. .
[32] Altschuler* JM, Talwar* K. Privacy of Noisy Stochastic Gradient Descent: More Iterations without More Privacy Loss. In: NeurIPS; 2022. .
[33] Bonawitz K, Ivanov V, Kreuter B, Marcedone A, McMahan HB, Patel S, et al. Practical secure aggregation for privacy-preserving machine learning. In: Proc. ACM SIGSAC Conf. Comput. Commun. Secur. (CCS). Dallas, TX; 2017. p. 1175-91.
[34] Bell JH, Bonawitz KA, Gascón A, Lepoint T, Raykova M. Secure Single-Server Aggregation with (Poly)Logarithmic Overhead. In: Proc. ACM SIGSAC Conf. Comput. Commun. Secur. (CCS). online; 2020. p. 1253-69.
[35] Kairouz P, Liu Z, Steinke T. The distributed discrete gaussian mechanism for federated learning with secure aggregation. In: International Conference on Machine Learning; 2021. p. 5201-12.
[36] Hälsodata som nationell resurs för framtidens hälso- och sjukvård;. Regeringskansliet, Sverige, S2022:04.
[37] Utredningen om hälsodata som nationellt intresse – en lagstiftning för interoperabilitet;. Regeringskansliet, Sverige, S2022:10.
[38] Uppdrag att genomföra kunskapshöjande insatser avseende integritets- och dataskyddsfrågor inom innovations-, utvecklings- och införandeprocesser;. Government of Sweden, N2020/01266.
[39] Kardeby V, Ardeshiri T, Eklund D. Bilaga nationellt datalagringsutrymme för bilddiagnostik. Sweden; 2022.
[40] Vinnova. [DNR I2021/02737] Slutrapport i regeringsuppdraget att kartlägga behov av utvecklingsinsatser för datadelning. Sweden; 2022.
[41] Decentralized AI;. AI Sweden. Accessed: 2023-03-30. https://www.ai.se/en/projects-9/decentralized-ai.
[42] Regulatory Sandboxes and Experimentation Clauses as tools for an innovation-friendly, future-proof and resilient regulatory framework that masters disruptive challenges in the digital age - Council conclusions, 16 November 2020.;. Accessed: 2023-03-30.
[43] IMY-2023-2602. Federerad maskininlärning mellan två vårdgivare. Slutrapport om Integritetsskyddsmyndighetens pilotprojekt med regulatorisk testverksamhet om dataskydd. Sweden; 2023.
[44] Framingham 10-year coronary heart disease risk. Kaggle;. Available from: https://www.kaggle.com/amanajmera1/framingham-heart-study-dataset/data.
[45] Tschandl P. The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. Harvard Dataverse; 2018. Available from: https://doi.org/10.7910/DVN/DBW86T.
[46] Yang J, Shi R, Ni B. MedMNIST Classification Decathlon: A Lightweight AutoML Benchmark for Medical Image Analysis. In: IEEE 18th International Symposium on Biomedical Imaging (ISBI); 2021. p. 191-5.
[47] Yang J, Shi R, Wei D, Liu Z, Zhao L, Ke B, et al. MedMNIST v2: A Large-Scale Lightweight Benchmark for 2D and 3D Biomedical Image Classification. arXiv preprint arXiv:211014795. 2021.
[48] Yousefpour A, et al. Opacus: User-Friendly Differential Privacy Library in PyTorch. arXiv preprint arXiv:210912298. 2021.
[49] He K, Zhang X, Ren S, Sun J. Deep Residual Learning for Image Recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE; 2016. .
[50] ImageNet;. http://www.image-net.org.
[51] Zhuang F, Qi Z, Duan K, Xi D, Zhu Y, Zhu H, et al. A Comprehensive Survey on Transfer Learning. Proceedings of the IEEE. 2021;109(1):43-76.

Preliminary Results on the use of Artificial Intelligence for Managing Customer Life Cycles

Jim Ahlstrand[1,2], Martin Boldt[1], Anton Borg[1] and Håkan Grahn[1]
{jim.ahlstrand, martin.boldt, anton.borg, hakan.grahn}@bth.se

***Abstract*— During the last decade we have witnessed how artificial intelligence (AI) have changed businesses all over the world. The customer life cycle framework is widely used in businesses and AI plays a role in each stage. However, implementing and generating value from AI in the customer life cycle is not always simple. When evaluating the AI against business impact and value it is critical to consider both the model performance and the policy outcome. Proper analysis of AI-derived policies must not be overlooked in order to ensure ethical and trustworthy AI. This paper presents a comprehensive analysis of the literature on AI in customer life cycles (CLV) from an industry perspective. The study included 31 of 224 analyzed peer-reviewed articles from Scopus search result. The results show a significant research gap regarding outcome evaluations of AI implementations in practice. This paper proposes that policy evaluation is an important tool in the AI pipeline and empathizes the significance of validating both policy outputs and outcomes to ensure reliable and trustworthy AI.**

***Index Terms*— artificial intelligence, customer life cycle, machine learning, policy evaluation**

I. Introduction

AI is having a significant impact on businesses worldwide. One example is AI ranking systems in the content creation industry. This adoption has resulted in a transformation of the way content is produced, with the goal of ranking higher by AI models and gaining a competitive advantage. This is a good example of implicit policy impact, in which the intended outcome, information ranking, represents just one of the actual outcomes. The effect of such changes are beyond the scope of this article, but it is critical to emphasize the difference of intended and actual policy outcomes.

AI is not a goal in itself; rather, it is a tool with the purpose of adding value. When businesses use AI to gain a competitive advantage, organizational efficiency, effectiveness, and customer relations, evaluating the model performance, i.e. policy output, is just as important as evaluating the actual policy outcome. Decisions made by AI should be held to the same standards as those made by their human counterparts.

As AI systems become more advanced and widely used in businesses, it is essential to ensure that they are making decisions that align with the values and goals of the organization, and that they are not causing unintended harm or bias. By holding AI systems to the same standards as human decision-makers, businesses can ensure that their AI policies and models are transparent, accountable, and trustworthy. This can help to build trust with customers and stakeholders, and ultimately lead to greater success and sustainability for the business.

In the telecom industry, the term "churn" is frequently used to describe when a consumer cancels a subscription. No matter how effective an AI model is at predicting churn, it does not, by itself, benefit the company. However, the model does add value when it is part of churn prevention efforts. The output, in this case, can be described as ranking customers by churn propensity, while the desired outcome is to lower the churn rate. Suppose a decision maker designs a policy where a discount is offered to the highest churn-risk customers each month. A model with a high recall but low precision may result in the treatment of customers who were not going to churn but now generate less revenue. In contrast, a model with high precision but low recall target only customers who will churn regardless of the discount, and the treatment has no effect. Even with perfect predictions, the sensitivity to proactive retention treatments must be taken into account to ensure the best possible outcome [1]. To avoid adversarial effects, the policy must be carefully considered and evaluated in relation to the desired outcome.

The use of AI to improve customer life cycles is not a novel concept. Churn prediction models, for example, have been around since the early 2000s [2]. Based on the results presented from the latest research, it is tempting to conclude that highly accurate propensity or ranking models imply higher business value. However, policies based on AI models may not produce the desired results in practice. The industry is eager to adopt AI, but coverage of quantitative evaluations of AI based policies are almost non-existent. It is necessary to understand the risks and impact of AI, wherever it is implemented and the customer life cycle is especially exposed as it often involves direct interaction with customers.

The objective of this study is to highlight the importance of policy evaluations in the context of customer life cycles utilizing AI and machine learning (ML). Through an analysis of recent literature, we seek to demonstrate the need for evaluating the effectiveness of AI-powered policies and decision support models. By emphasizing the significance of policy evaluations, this study aims to promote the development of more reliable and trustworthy AI practices in customer life cycle management. Hence, this study aims to answer the following research questions:

RQ1: What conclusions can be drawn from the literature regarding how AI is applied throughout the customer life cycle?

*This work was funded by Telenor Sverige AB.
[1] Blekinge Institute of Technology, SE-371 79 Karlskrona, Sweden.
[2] Telenor Sverige AB, SE-371 80 Karlskrona, Sweden.

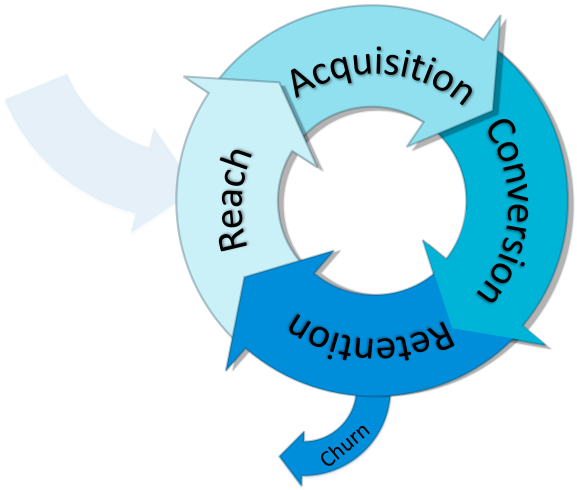

Fig. 1: Customer life cycle stages

RQ2: To what extent have policies derived from AI models been evaluated in customer life cycles?

This article argues that AI has the ability to empower customer life cycle management, but its successful implementation requires a careful balance between automation and human-in-the-loop, as well as a commitment to ethical AI practices. The main contributions in this paper are the following:

- First, we propose the use of policy evaluation to assess the impact of AI on the customer life cycle.
- Second, from an industrial perspective, a research gap is identified and motivated as a direction for future research.
- Third, this paper provides a high-level overview of how AI is applied and how its impact is measured throughout the customer life cycle.

II. Background

Customer relationship management (CRM) focuses on building and maintaining relationships with customers to maximizing their lifetime value. It is a term that encompasses cross-functional properties such as multichannel integration, customer interaction, data collection and analytics [3]. Customer life cycle management (CLM) is a broader approach, concerned with managing the entire customer life cycle to deliver the overall best possible customer experience.

A. Customer life cycle

The definitions of what stages should be included in the customer life cycle have varied over time. However, there are a few stages that are consistently covered in the most recent research Figure 1.

Reach is about brand awareness and equity. Making sure all potential customers are aware of the products and services offered, and what solutions the company provides to solve their problems. Customer profiling, segmentation, and targeting fall under this category [4].

Acquisition is when the customer takes action to engage in communicating and building a relation. How this looks depends on the channel, e.g. visiting the company website or calling a sales representative is handled differently. Product recommendations, personalization, as well as lead ranking are applied in this stage [4].

Conversion happens when the prospect finalizes a purchase and becomes a value-added customer [4]. Dynamic pricing and content curation based on the buyers journey are two examples of conversion improvements.

Retention, loyalty or advocacy is where you want the customer to remain. Retaining a customer is often many times cheaper than acquiring new ones. Up- and cross-sell, customer support, and co-creation are common activities here. Satisfied and loyal customers become brand advocates, expanding the reach and completing the circle.

Churn is defined as the percentage of customers that decide to exit, i.e, the customer stops buying products or services from a company [5]. In this analysis, churn is considered a separate stage in the life cycle as it also comes with its separate actions, e.g. trusted advisor, personalized offers, exit management, and re-targeting previous customers.

B. Policy evaluation

A policy evaluation is the process of objective, systematic, and empirical review of the effects a policy has on its target. How it is performed depends on the policy and the result it tries to achieve.

Formative analysis assures that a policy is feasible and appropriate before it is implemented fully. This could be a conceptual framework from which conclusions about the outcome are inferred.

Process evaluation makes sure that the policy has been implemented correctly, activities are carried out, and are reaching the targeted population as intended. Process evaluation can highlight implementation issues early but does not provide insights into the effects of the policy.

Cost-benefit analysis sums the expected future reward of an action and subtracts the expected cost. This is rather straight-forward when the goal is financial and is commonly used in businesses. However, it may be misleading when the benefits are partially or entirely intangible.

Impact analysis is an evidence-based method that tries to answer what the effects are of an action compared to inaction. Causal relationships have to be defined and tested to answer what outcomes are directly attributable to the policy.

Qualitative evaluation is useful to determine effects on opinions, attitudes, motivations or experiences. Usually, it takes the form of surveys, questionnaires, focus groups or interviews.

III. Methodology

To answer the research questions, a review of the literature was conducted. Scopus was chosen as the bibliographic database due to its extensive coverage of peer-reviewed papers in data science. The search query was developed with the intention of achieving high recall and capturing all stages of the customer life cycle. It also emphasizes AI implementations and empirical findings. The query's first component contains keywords related to the customer life cycle, these were selected based on a literature review [4] as well as gray literature [6], [7]. The second section contains keywords related to AI and relevant subdomains. The query, when executed (January 27, 2023), returned 224 results.

```
TITLE-ABS-KEY(customer w/5 (lead OR prospect OR
    reach OR acquisition OR conversion OR retention
     OR churn OR "life cycle" OR relationship OR
    experience OR journey) AND ("artificial
    intelligence" OR "machine learning" OR "big
    data" OR "expert system" OR "deep learning" OR
    ai OR ml OR dl) AND empirical*)
```

Each result was screened based on title and abstract using the following criteria:

- Must cover one or more stages of the customer life cycle,
- The article must be peer-reviewed with available, non-retracted, full-text in English,
- The article must be published in a journal, workshop, or conference,
- Implements AI or subdomains, e.g. ML, in one or more stages in the customer life cycle,
- Presents empirical evidence, excluding surveys, questionnaires, and interviews[1],
- Excluding conceptual or theoretical papers, e.g. comparing or evaluating the performance of specific AI implementations, data collection or training.

Figure 2 shows how the screening and analysis process was conducted. The titles and abstracts were read and compared to the above-mentioned inclusion criteria. The full text was obtained if the abstract did not reliably exclude the paper. This stepädecreased the number of publications from 224 to 48. For the remaining articles, the complete text was reviewed and assessed again using the inclusion criteria and research questions. Following this stage, the final set contains 31 relevant articles.

IV. Results

Table I shows the distribution of journals and conferences per article, as well as figure 3 which shows the distribution per year. Table II shows the method applied per stage in the customer life cycle. Only five of the 31 articles included any type of policy evaluation. These includes formative analysis based on simulations (P_t), and more practical methods such as randomized tests (P_p). This is a clear indication that future research must focus on theoretical and practical policy evaluations.

[1]Surveys are excluded since they focus on subjective experiences rather than objective outcomes, which makes them ineffective for answering the research questions. Surveys may also be subject to sampling errors, measurement bias, and high cost due to the manual handling [8]

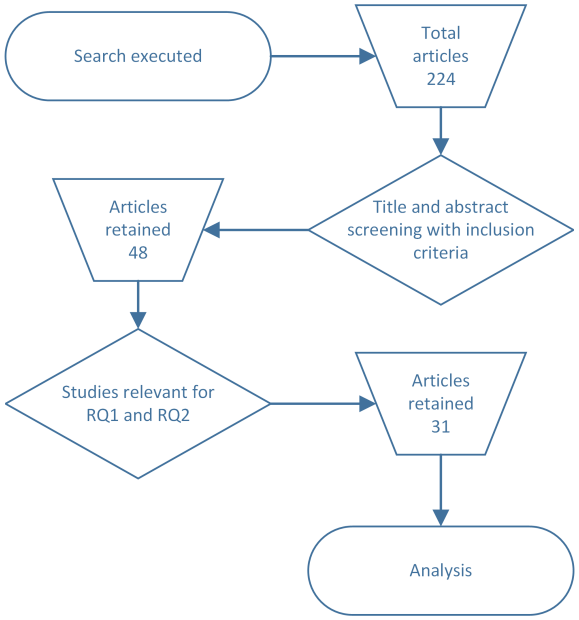

Fig. 2: Methodology flowchart

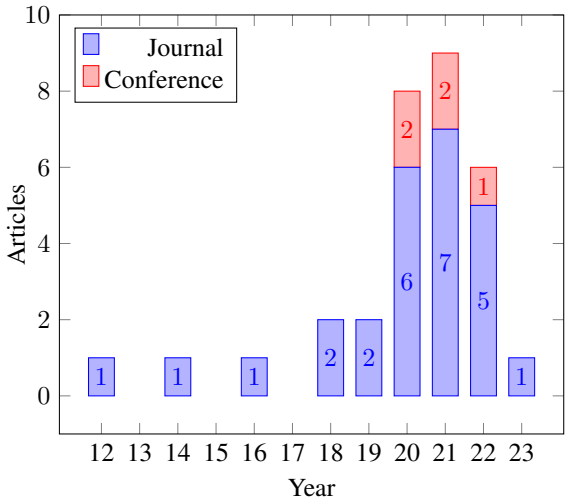

Fig. 3: Type of document per for each year in the range 2012 and 2023.

A. Reach & Acquisition

The two first stages of the customer life cycle are reach and acquisition. The distinction between the two are described in the background in Section II-A, however, in practice there are overlaps. The literature usually look at the two stages combined when implementing AI, for instance customer targeting and customer response rate [9], or early customer life time value forecasting [10].

Haupt and Lessmann used statistical models and ML to improve the targeting policy of an e-coupon campaign [9]. The authors analyzes the targeting decision problem and argues that the treatment cost is not fixed in practice but is dependant on the customer response probability. The derived

TABLE I: Number of articles per source

Publisher	Source	Articles
ACM	ACM Transactions on Knowledge Discovery from Data	1
ACM	ICMI Companion - Companion Publ. Int. Conf. Multimodal Interact.	1
Elsevier	Applied Energy	1
Elsevier	European Journal of Operational Research	1
Elsevier	Expert Systems with Applications	2
Elsevier	International Journal of Information Management	1
Elsevier	Applied Soft Computing Journal	1
Elsevier	Knowledge-Based Systems	1
Emerald	Asia Pacific Journal of Marketing and Logistics	1
Emerald	Information Technology and People	1
Emerald	Journal of Enterprise Information Management	1
Emerald	Journal of Service Management	1
Emerald	Kybernetes	1
Hindawi	Computational Intelligence and Neuroscience	1
Hindawi	Discrete Dynamics in Nature and Society	1
IEEE	2020 IEEE 5th International Conference on Cloud Computing and Big Data Analytics	1
IEEE	2020 IEEE International Conference on Big Data	1
IGI Global	Journal of Global Information Management	1
MDPI	Future Internet	1
MDPI	Risks	1
MDPI	Sensors	1
SAGE	Journal of Marketing Research	2
Springer	Annals of Operations Research	1
Springer	Computational Economics	1
Springer	Lecture Notes in Computer Science	2
SSRG	International Journal of Engineering Trends and Technology	1
Taylor & Francis	Journal of Management Information Systems	1
Wiley	International Journal of Intelligent Systems	1
Total		31

TABLE II: Method per customer life cycle stage. St = statistical methods, ML = machine learning, DL = deep learning, P_t = policy tested in theory, and P_p = policy tested in practice.

Stage	St	ML	DL	Total	P_t	P_p
Reach	1	4		5	1	
Acquisition		2		3		
Conversion		4	1	4		
Retention	5	3		8		2
Churn	2	9		11	2	
Total	8	22	1	31	3	2

policy was tested by simulating the stochastic variables based on real world data and the result was an increase in the overall campaign profit.

Traditional marketing channels are typically comprised of broadcasts; these channels have the potential to reach a large number of customers. However, compared to unicast channels, broadcasts are expensive and not very efficient. The problem with unicast channels is to find and target the right customers. Customer segmentation can improve the effectiveness of targeted marketing. The assumption is that there exists groups of customers that have similar characteristics and therefore are more likely to purchase similar products. Several studies have shown varying success at implementing AI for this purpose [11]–[14]. In practice, this usually means that the segmentation clusters are manually labeled and paired with a marketing campaign based on perceived personas. However, little is known about the actual outcomes of customer segmentation-based targeting or other uses of customer segmentation.

Customer lifetime value (CLV) can drive targeting policies to focus resources on the most valuable clients and therefor also competitive advantage and profit growth [15]. CLV is typically based on demographics and historical purchase data. Forecasting CLV based on historical purchase data only works after you have gathered enough data about a specific customers' spending pattern. This inability to make inferences early is often called a "cold start" or "bootstrap" problem. Padilla and Ascarza [10] developed a modeling framework called First-Impression Model (FIM) which can identify high-value customers at the acquisition stage and thereof, alleviating the cold start problem of CRM. The authors evaluate their model by running simulations based on real data and show promising empirical results. They discuss policies that could be implemented in practice for e.g. customer targeting based on the predictions of FIM. However, the evaluation of real-world policies were left to future research.

B. Conversion

Conversion is the stage where a prospect becomes a value-added customer. A strong application of AI within conversion is dynamic pricing. The desired output of such models are to predict the price for which the customer most likely will convert. Pricing strategies are typically developed through supply and demand modeling, in which both supply and demand participate in moving the price one way or another [16]. In some studies the supply side is known or presumed fixed, which lets the AI model focus on predicting the demand [17] or vice versa. However, the forecast would not by itself affect the desired outcome which is increased profit. When defining a pricing strategy based on forecasting demand, supply or both; it is important to factor the risks

of setting the price too low, or too high. The policy should be evaluated based on the desired outcome e.g. increased profits.

Some studies examined user behavior from various touchpoints. If touchpoints can adapt to individual user behavior, conversion rates may increase. In practice, this usually entails generalizing across a population in order to infer individual behavior patterns [18]. Alfian et al. [19] proposes a statistical model for extracting association rules from real-time online transactional data. The model was theoretically evaluated by the authors in a simulated environment. In which the model successfully predicted customer behavior and increased sales of recommended products.

Recommendation systems were initially developed to improve the information retrieval process. From a business perspective this usually means assisting the customers in finding the most relevant products. These may manifest as search intent models [20] or dynamic content curation based on the customers buying cycle. However, the recommendations could just as well have the opposite effect if the choice of recommendation engine does not fit the purpose or is badly implemented [21].

C. Retention

When a prospect converts to a value added customer they enter the retention stage. Many activities go into customer retention, and success in retaining customers can be measured by the churn rate. As stated in the background II-A, churn is treated as a separate stage from retention activities in this analysis.

Churn prevention is executed when there are clear indications of churn, whereas retention activities focus on building loyalty and a healthy relationship with the customer. Hedonic aspects, such asăcustomer experience and satisfaction,ăare important for a healthy customerărelationship. Customer loyalty however, requires a business to consistently delivery a positiveăexperience [22]. Customer service and support play a critical role in ensuring that problems voiced by customers are resolved in a timely and satisfactory manner. Customer loyalty may even increase after reporting a problem, if managed properly [23].

1) Experience & Satisfaction: Experience and satisfaction are difficult to quantify because they are subjective. Customers may find it difficult to express their experience due to the subjective properties. Customer experience (CX) measurement are however essential to businesses to ensure that customer journeys align with the company's vision and enable long-term trust and co-creation. Understanding the customer's emotions are key to also understand their behaviour [24].

Sidaoui, Jaakkola and Burton [25] proposes a chat bot framework that is based around "narrative inquiry via data collection mechanisms such as storytelling and interviews" to improve primary experience gathering. Deng and Murari [26] suggests that Crowdsourced Voice Feedback is a more reliable, customer-centric, less expensive and instantaneous method of measuring CX when using Intelligent Voice Assistants (IVA). With the use of ML based causal inference methods, they discovered that prompting customers for feedback had no negative impact on CX; however, when the elicitation frequency increased the response rate decreased. The authors also investigated the causality between response rate and types of questions asked and when they were asked. Lee, Tse, Zhang and Ma [27] extracted insights from customer reviews of short-term homestays and experiences using statistical methods. In theory, the method they present can provide hosts with valuable customer-centric and immediate insights, improving product quality and customer experience. However, the authors do not address potential risks or consequences in an online real-world setting.

Yu et al. [28] developed a decision system based on ML to optimize the delivery scheduling of food deliveries. The system controls the order delay when the probability of grouping orders of similar destinations are high, thus increasing the delivery efficiency. It is at the same time constrained by the probability of overtime penalties and average delivery time. The resulted policy was tested using an online A/B test in several cities. The decision system was compared to a fixed 2 minute delay policy and showed an improved grouping success of 41.20% compared to 22.19%. The article did not discuss explainability aspects or potential risks with variance or bias towards some types of orders compared to others.

2) Service & Support: Andrade and Moazeni describe a statistical model that predicts interactive voice response (IVR) transfer rates with an area under the curve between 77% and 95% depending on the callers location [29]. They present theoretical impacts, e.g. the model could improve customer satisfaction by bypassing the self service IVR solution, but does not evaluate any practical policy based on this assumption. It could be argued that a simpler policy for achieving this goal is to make the self service IVR optional, considering that the authors also stated that "dealing with automated customer service platforms before reaching to an agent is considered to be the most frustrating part of a poor contact experience".

Being able to forecast obsolescence enables a business to better allocate resources for e.g. stocking up on parts and repairs. The idea is that being proactive in this area allows the business to cut costs by improving the organizational efficiency, and at the same time provide a better customer experience. AI can be employed to predict the obsolescence date of products based on individual wear and tear [30]. Little is known about the magnitude of the impact by forecasting obsolescence to prioritize preventative actions and avoid costs.

Failures will occur despite how diligently a company tries to avoid them. There is a constant balance between risk mitigation and cost containment. There are factors that influence this balance in both directions, and it all boils down to how much the customer is willing to pay. Kim et al. [31] analyze these factors in detail and how they impact willingness to pay (WTP). The authors demonstrate that ML can be used to approximate the customer damage function

(CDF) and discuss how it can be used in practice to improve the product offerings. Future research will show how such policies manifests in the real world.

D. Churn

Customer churn is defined as the percentage of customers that decide to exit, i.e the customer stops buying products or services from a company. All businesses eventually reach a point where market saturation makes acquiring new customers more expensive than retaining existing customers [5], [32]. Businesses operating in new markets or startups place less emphasis on churn rate while growth is still cheap. Churn metrics also indicates the overall satisfaction and willingness to do repeated business. Churn rate is perhaps most notable in companies that offer services or subscriptions, e.g. telecommunications, cloud or streaming media companies. In the literature there were mainly two forms of churn modelling presented, prediction and prevention models. However, most research focus on the churn prediction problem and little is known about the effects of churn prevention policies.

1) Prediction: Forecasting churn can be very effective, but the data is difficult to collect and the available data is naturally very imbalanced as the churn rate preferably should be as low as possible [32], [33]. It is also difficult to capture all the relevant features as both internal and external events can affect the reason to churn [34], [35]. There appears to be no widely used benchmark dataset for this purpose, making it difficult to compare literature. Nonetheless, the churn prediction problem has received considerable attention [36]–[40].

2) Prevention: Based on churn predictions, policies can be formed to reduce the churn rate, i.e. churn prevention. This step is often overlooked by the available literature. It is not enough to present a model with high evaluation scores for it to be useful in practice. The model must support a policy that can identify the right customers and the right treatment to lower the churn rate. Customer respond differently to different treatment, therefore, individual treatment effect (ITE) is an essential driver in the targeting decision to ensure an optimal outcome. Using the naive approach and targeting high risk churners while disregarding ITE may even have the opposite effect [1]. The possibility to alter the churn prediction optimizer for either precision or recall enables managers to better tailor the model to better suit their specific targeting goals [5]. There are trade-offs when using models based on imperfect information[2] that needs to be acknowledged during the targeting decision.

V. Discussion

Customer life cycles are broad and touch on so many different domains that it is difficult to capture all of them in a single review. Even though the database query was designed for recall it may return more results if each stage were queried separately, although this would presumably also increase the required effort to review greatly. The authors acknowledge that the query is limited to articles using the term "empirical*" that may not fully represent the entire set of empirical work. This paper is also limited to business applications of AI.

Ranking models have become a very popular tool for improving information retrieval and recommendations. In a digital age where data is abundant, the ability to quickly and effectively retrieve relevant data is demonstrated at several stages in the customer life cycle. It provides customer targeting and segmentation for improving reach and acquisition. Dynamically curating web pages and recommending relevant products provide increased conversion. The ranking may improve retention and loyalty by aiding customer service in retrieving relevant information to solve problems quicker. However, it may also target consumers who aren't qualified as long-term loyal customers or offer solutions that don't genuinely fix their problem, which has a negative impact on customer relations. A system that automatically gives out discounts based on a customer's churn probability score may encourage customers to churn more frequently in the long run. These examples demonstrate the effects a flawed policy could expose in a real-world setting.

Customer acquisition cost, customer satisfaction, and lifetime value are rarely discussed alongside propensity to buy, customer segmentation, and targeting models. This may not be a problem if the company's only goal is to increase sales or market share. But in reality, it is much more likely that customers are valued different for strategic and long-term growth purposes. It may be reasonable to take a chance on a high-value customer if the treatment cost is not prohibitively costly in relation to the conversion probability, even if that probability is lower than that of the alternative low-value customers. There may also be ethical considerations that influence whether or not to associate with certain customers over others, such as when the customer's reputation conflicts with the company's values. The desired outcome must be the primary consideration when deciding on the metrics of success to use for AI implementations. These trade-offs are not discussed to any greater extent in the literature.

Churn prevention as a policy based on churn prediction appears to be mostly overlooked by the literature. Consider the combination of a churn prediction model, which feeds a treatment response model to determine the most effective treatment, which in turn feeds a treatment cost prediction model. This problem is similar to the e-coupon targeting presented by Haupt and Lessmann as described in section IV-A.

There are several ways to evaluate policies, as described in section II-B. For churn prevention, a formative analysis could be done using historical data to identify patterns and trends in customer behavior that may be contributing to churn. Cost-benefit analysis may prove to be complex, as the cost of the treatment can vary based on the accuracy of the prediction model, as inaccurate predictions may lead to incorrect treatments, resulting in higher costs. Accurately assessing the costs and benefits of an intervention requires consideration of both the treatment cost and the response

[2]The assumption that there will always be imperfect information about churn determinants is reasonable as the decision is influenced by both external and internal factors.

rate which are both dependent on the accuracy of the model. The gold standard for impact analysis is A/B testing which can identify which policies are most effective at achieving their desired outcomes, while also mitigating potential harms or unintended consequences. A/B testing does require interventions that entails risks which may not be acceptable in practice, for those cases one could look at e.g. double robust learners which works on observational data. Finally, a qualitative assessment could include methods such as focus groups, interviews, surveys, and net promoter score (NPS) to gather feedback on the experience of an AI-powered system and how it affected the decision-making process.

The choice of optimization is imperative to achieve desired outcomes. When employing AI to automatically elicit customer feedback [26] or mining insights from product reviews [41], the policy output is *when to trigger elicitation* and *the similarity score of product reviews* respectively. However, the desired policy outcome is an improved customer agility and experience. The effectiveness of this outcome in practice depends on what the AI is optimized for. In order to obtain accurate insights, the data must be representative of the population. Regarding feedback elicitation, optimizing for response rate may have a negative impact on the total number of responses as only the customers that are guaranteed to respond will be elicited. On the other hand, optimizing on the number of responses may elicit feedback too often and reduce both the response rate and customer experience. Zhou et al. demonstrated that there are both positive and negative impacts on product performance depending on how the reviews are used. Some products do benefit from high agility, while others may suffer from rapidly changing behavior. Only optimizing for the agility metric may not result in deliveries that actually improve the customer experience and product performance.

VI. Conclusions and future work

In summary, AI implementations have been mapped to each of the customer life cycle stages and thoroughly analyzed. Theoretical and practical implications of AI were discussed for each stage. AI model performance evaluations were frequent, however, the model performance does not necessarily translate to the expected real-world outcomes. The evidence for successful and acceptable policy outcomes were almost non-existent. Successful customer life cycle management takes a broad view and responsibility for the entire customer experience. The effects of AI on the overall experience were not presented in the literature. This concludes the answer to RQ1.

Only five papers that studied the outcomes of AI policies were found in this analyzis. Three of which were formative [1], [5], [9] and the other two performed an online impact analysis [28] and cost-benefit analysis [26]. For industries looking at employing AI in their business, policy evaluations supports the business case and trust. This review highlights a research gap when it comes to evaluating the outcomes of AI in customer life cycle management. Policy evaluations are helpful in assuring stakeholders that the AI not only performs well on paper, but also works as intended, with acceptable outcomes and with no adversarial biases or ethical concerns. The significance of human-centered evaluative research is critical to business ethics and society as AI becomes more prevalent in real-world settings. This concludes the answer to RQ2.

Future work can branch into several directions. First, there is a need to establish an unambiguous and easy-to-follow framework for how a business should perform policy evaluations on AI implementations. What are the steps, necessary tools, metrics, terminology, and processes involved? How can policy evaluations be part of the AI pipeline from start to production?

Second, guidelines for trustworthy and ethical AI in businesses should be included in the policy evaluation process.Transparency is key to growing trust, but further research is necessary to understand how transparency is best communicated. In what ways should a business expose that a decision was made by AI? Is it possible to challenge the decisions? In what ways can a customer voice concerns or problems they experience from AI decisions?

A set of standard datasets for evaluation is required to further enrich the literature on AI in customer life cycles. To my knowledge, no widespread standard evaluation frameworks exist for any of the steps in the customer life cycle, such as customer segmentation, dynamic pricing, or churn prevention.

References

[1] E. Ascarza, "Retention futility: Targeting high-risk customers might be ineffective," *J. Mark. Res.*, vol. 55, no. 1, pp. 80–98, 2018, Publisher: American Marketing Association, ISSN: 00222437 (ISSN). DOI: 10.1509/jmr.16.0163.

[2] G. Madden, S. J. Savage, and G. Coble-Neal, "Subscriber churn in the australian isp market," English, *Information Economics and Policy*, vol. 11, no. 2, pp. 195–207, 1999, Cited by: 49; All Open Access, Green Open Access, ISSN: 01676245. DOI: 10.1016/S0167-6245(99)00015-3. [Online]. Available: https://www.scopus.com/inward/record.uri?eid=2-s2.0-0032863138&doi=10.1016%2fS0167-6245%2899%2900015-3&partnerID=40&md5=da76c1adb611b4c7e07b0ca2f4fa0270.

[3] A. Payne and P. Frow, "A strategic framework for customer relationship management," *Journal of Marketing*, vol. 69, no. 4, pp. 167–176, 2005. DOI: 10.1509/jmkg.2005.69.4.167. eprint: https://doi.org/10.1509/jmkg.2005.69.4.167. [Online]. Available: https://doi.org/10.1509/jmkg.2005.69.4.167.

[4] M. Moradi and M. Dass, "Applications of artificial intelligence in b2b marketing: Challenges and future directions," *Industrial Marketing Management*, vol. 107, pp. 300–314, Nov. 1, 2022, ISSN: 0019-8501.

DOI: 10.1016/j.indmarman.2022.10.016. (visited on 01/20/2023).

[5] J. Kozak, K. Kania, P. Juszczuk, and M. Mitrga, "Swarm intelligence goal-oriented approach to data-driven innovation in customer churn management," *Int J Inf Manage*, vol. 60, 2021, Publisher: Elsevier Ltd, ISSN: 02684012 (ISSN). DOI: 10.1016/j.ijinfomgt.2021.102357.

[6] "Everything you need to know about customer lifecycle management." (Jan. 24, 2022), [Online]. Available: https://blog.hubspot.com/service/customer-lifecycle-management (visited on 03/16/2023).

[7] "Customer lifecycle management (CLM): The ultimate guide forbes advisor." (), [Online]. Available: https://www.forbes.com/advisor/business/customer-lifecycle-management/ (visited on 03/16/2023).

[8] R. M. Groves, *Survey errors and survey costs*. Wiley, 2004, ISBN: 978-0-471-67851-9.

[9] J. Haupt and S. Lessmann, "Targeting customers under response-dependent costs," *Eur J Oper Res*, vol. 297, no. 1, pp. 369–379, 2022, Publisher: Elsevier B.V., ISSN: 03772217 (ISSN). DOI: 10.1016/j.ejor.2021.05.045.

[10] N. Padilla and E. Ascarza, "Overcoming the cold start problem of customer relationship management using a probabilistic machine learning approach," *J. Mark. Res.*, vol. 58, no. 5, pp. 981–1006, 2021, Publisher: SAGE Publications Ltd, ISSN: 00222437 (ISSN). DOI: 10.1177/00222437211032938.

[11] P. Sarvari, A. Ustundag, and H. Takci, "Performance evaluation of different customer segmentation approaches based on RFM and demographics analysis," *Kybernetes*, vol. 45, no. 7, pp. 1129–1157, 2016, Publisher: Emerald Group Publishing Ltd., ISSN: 0368492X (ISSN). DOI: 10.1108/K-07-2015-0180.

[12] R. Mancisidor, M. Kampffmeyer, K. Aas, and R. Jenssen, "Learning latent representations of bank customers with the variational autoencoder," *Expert Sys Appl*, vol. 164, 2021, Publisher: Elsevier Ltd, ISSN: 09574174 (ISSN). DOI: 10.1016/j.eswa.2020.114020.

[13] M. Hossain, M. Sebestyen, D. Mayank, O. Ardakanian, and H. Khazaei, "Large-scale data-driven segmentation of banking customers," in *Proc. - IEEE Int. Conf. Big Data, Big Data*, Journal Abbreviation: Proc. - IEEE Int. Conf. Big Data, Big Data, Institute of Electrical and Electronics Engineers Inc., 2020, pp. 4392–4401, ISBN: 978-172816251-5 (ISBN). DOI: 10.1109/BigData50022.2020.9378483.

[14] A. Gramegna and P. Giudici, "Why to buy insurance? an explainable artificial intelligence approach," *Risks*, vol. 8, no. 4, pp. 1–9, 2020, Publisher: MDPI AG, ISSN: 22279091 (ISSN). DOI: 10.3390/risks8040137.

[15] J. Bauer and D. Jannach, "Improved customer lifetime value prediction with sequence-to-sequence learning and feature-based models," *ACM Trans. Knowl. Discov. Data*, vol. 15, no. 5, 2021, Publisher: Association for Computing Machinery, ISSN: 15564681 (ISSN). DOI: 10.1145/3441444.

[16] L. Li, S. Ma, X. Han, C. Zheng, and D. Wang, "Data-driven online service supply chain: A demand-side and supply-side perspective," *J. Enterp. Inf. Manage.*, vol. 34, no. 1, pp. 365–381, 2020, Publisher: Emerald Group Holdings Ltd., ISSN: 17410398 (ISSN). DOI: 10.1108/JEIM-11-2019-0352.

[17] M. Alauddin and C.-Y. Ting, "Digital click stream data for airline seat sale prediction using GBT," *Int. J. Eng. Trends Technol.*, no. 1, pp. 24–31, 2020, Publisher: Seventh Sense Research Group, ISSN: 23490918 (ISSN). DOI: 10.14445/22315381/CATI3P204.

[18] M. Musaolu, M. Bekler, H. Budak, C. Akçelik, and M. Aktas, "On theămachine learning based business workflows extracting knowledge fromălarge scale graph data," in *Lect. Notes Comput. Sci.*, Journal Abbreviation: Lect. Notes Comput. Sci., vol. 13381 LNCS, Springer Science and Business Media Deutschland GmbH, 2022, pp. 463–475, ISBN: 03029743 (ISSN); 978-303110547-0 (ISBN). DOI: 10.1007/978-3-031-10548-7_34.

[19] G. Alfian, M. Ijaz, M. Syafrudin, M. Syaekhoni, N. Fitriyani, and J. Rhee, "Customer behavior analysis using real-time data processing: A case study of digital signage-based online stores," *Asia Pac. J. Mark. Logist.*, vol. 31, no. 1, pp. 265–290, 2019, Publisher: Emerald Group Holdings Ltd., ISSN: 13555855 (ISSN). DOI: 10.1108/APJML-03-2018-0088.

[20] J. Ma, X. Guo, and X. Zhao, "Identifying purchase intention through deep learning: Analyzing the q &d text of an e-commerce platform," *Ann. Oper. Res.*, 2022, Publisher: Springer, ISSN: 02545330 (ISSN). DOI: 10.1007/s10479-022-04834-w.

[21] M. Gorgoglione, U. Panniello, and A. Tuzhilin, "Recommendation strategies in personalization applications," *Inf Manage*, vol. 56, no. 6, 2019, Publisher: Elsevier B.V., ISSN: 03787206 (ISSN). DOI: 10.1016/j.im.2019.01.005.

[22] S. Hyken, *The cult of the customer: Create an amazing customer experience that turns satisfied customers into customer evangelists*. Sound Wisdom, 2020, ISBN: 978-1640951532.

[23] D. Research, *Quantifying the business impact of customer service in australia*, Apr. 2019. [Online]. Available: https://zen-marketing-content.s3.amazonaws.com/content/Zendesk2019_Quantifying_the_business_impact_of_customer_service_Australia.pdf (visited on 03/03/2023).

[24] M. Lee, S. Lee, and Y. Koh, "Multisensory experience for enhancing hotel guest experience: Empirical evi-

dence from big data analytics," *Int. J. Contemp. Hosp. Manage.*, vol. 31, no. 11, pp. 4313–4337, 2019, Publisher: Emerald Group Holdings Ltd., ISSN: 09596119 (ISSN). DOI: 10.1108/IJCHM-03-2018-0263.

[25] K. Sidaoui, M. Jaakkola, and J. Burton, "AI feel you: Customer experience assessment via chatbot interviews," *J. Serv. Manage.*, vol. 31, no. 4, pp. 745–766, 2020, Publisher: Emerald Group Holdings Ltd., ISSN: 17575818 (ISSN). DOI: 10.1108/JOSM-11-2019-0341.

[26] Y. Deng and S. Murari, "When a voice assistant asks for feedback: An empirical study on customer experience with a/b testing and causal inference methods," in *ICMI Companion - Companion Publ. Int. Conf. Multimodal Interact.*, Journal Abbreviation: ICMI Companion - Companion Publ. Int. Conf. Multimodal Interact., Association for Computing Machinery, Inc, 2021, pp. 183–191, ISBN: 978-145038471-1 (ISBN). DOI: 10.1145/3461615.3485403.

[27] C. Lee, Y. Tse, M. Zhang, and J. Ma, "Analysing online reviews to investigate customer behaviour in the sharing economy: The case of airbnb," *Inf. Technol. People*, vol. 33, no. 3, pp. 945–961, 2020, Publisher: Emerald Group Holdings Ltd., ISSN: 09593845 (ISSN). DOI: 10.1108/ITP-10-2018-0475.

[28] Y. Yu, Q. Zhou, S. Yi, *et al.*, "Delay to group in food delivery system: A prediction approach," in *Lect. Notes Comput. Sci.*, Journal Abbreviation: Lect. Notes Comput. Sci., vol. 12837 LNCS, Springer Science and Business Media Deutschland GmbH, 2021, pp. 540–551, ISBN: 03029743 (ISSN); 978-303084528-5 (ISBN). DOI: 10.1007/978-3-030-84529-2_46.

[29] R. Andrade and S. Moazeni, "Transfer rate prediction at self-service customer support platforms in insurance contact centers," *Expert Sys Appl*, vol. 212, 2023, Publisher: Elsevier Ltd, ISSN: 09574174 (ISSN). DOI: 10.1016/j.eswa.2022.118701.

[30] K.-S. Moon, H. Lee, H. Kim, H. Kim, J. Kang, and W. Paik, "Forecasting obsolescence of components by using a clustering-based hybrid machine-learning algorithm," *Sensors*, vol. 22, no. 9, 2022, Publisher: MDPI, ISSN: 14248220 (ISSN). DOI: 10.3390/s22093244.

[31] M. Kim, B. Lee, H. Lee, S. Lee, J. Lee, and W. Kim, "Robust estimation of outage costs in south korea using a machine learning technique: Bayesian tobit quantile regression," *Appl. Energy*, vol. 278, 2020, Publisher: Elsevier Ltd, ISSN: 03062619 (ISSN). DOI: 10.1016/j.apenergy.2020.115702.

[32] H. Thakkar, A. Desai, S. Ghosh, P. Singh, and G. Sharma, "Clairvoyant: AdaBoost with cost-enabled cost-sensitive classifier for customer churn prediction," *Comput. Intell. Neurosci.*, vol. 2022, 2022, Publisher: Hindawi Limited, ISSN: 16875265 (ISSN). DOI: 10.1155/2022/9028580.

[33] P. imovi, C. Chen, and E. Sun, "Classifying the variety of customers online engagement for churn prediction with a mixed-penalty logistic regression," *Comput. Econ.*, 2022, Publisher: Springer, ISSN: 09277099 (ISSN). DOI: 10.1007/s10614-022-10275-1.

[34] L. Almuqren, F. Alrayes, and A. Cristea, "An empirical study on customer churn behaviours prediction using arabic twitter mining approach," *Future Internet*, vol. 13, no. 7, 2021, Publisher: MDPI AG, ISSN: 19995903 (ISSN). DOI: 10.3390/fi13070175.

[35] X. Zhang, J. Zhu, S. Xu, and Y. Wan, "Predicting customer churn through interpersonal influence," *Knowl Based Syst*, vol. 28, pp. 97–104, 2012, ISSN: 09507051 (ISSN). DOI: 10.1016/j.knosys.2011.12.005.

[36] M. Farquad, V. Ravi, and S. Raju, "Churn prediction using comprehensible support vector machine: An analytical CRM application," *Appl. Soft Comput. J.*, vol. 19, pp. 31–40, 2014, ISSN: 15684946 (ISSN). DOI: 10.1016/j.asoc.2014.01.031.

[37] M. Li, C. Yan, W. Liu, and X. Liu, "An early warning model for customer churn prediction in telecommunication sector based on improved bat algorithm to optimize ELM," *Int J Intell Syst*, vol. 36, no. 7, pp. 3401–3428, 2021, Publisher: John Wiley and Sons Ltd, ISSN: 08848173 (ISSN). DOI: 10.1002/int.22421.

[38] X. Hu, Y. Yang, L. Chen, and S. Zhu, "Research on a customer churn combination prediction model based on decision tree and neural network," in *IEEE Int. Conf. Cloud Comput. Big Data Anal., ICCCBDA*, Journal Abbreviation: IEEE Int. Conf. Cloud Comput. Big Data Anal., ICCCBDA, Institute of Electrical and Electronics Engineers Inc., 2020, pp. 129–132, ISBN: 978-172816024-5 (ISBN). DOI: 10.1109/ICCCBDA49378.2020.9095611.

[39] L. Cheng, C.-C. Wu, and C.-Y. Chen, "Behavior analysis of customer churn for a customer relationship system: An empirical case study," *J. Global Inf. Manage.*, vol. 27, no. 1, pp. 111–127, 2019, Publisher: IGI Global, ISSN: 10627375 (ISSN). DOI: 10.4018/JGIM.2019010106.

[40] M. Zhao, Q. Zeng, M. Chang, Q. Tong, and J. Su, "A prediction model of customer churn considering customer value: An empirical research of telecom industry in china," *Discrete Dyn. Nat. Soc.*, vol. 2021, 2021, Publisher: Hindawi Limited, ISSN: 10260226 (ISSN). DOI: 10.1155/2021/7160527.

[41] S. Zhou, Z. Qiao, Q. Du, G. Wang, W. Fan, and X. Yan, "Measuring customer agility from online reviews using big data text analytics," *J Manage Inf Syst*, vol. 35, no. 2, pp. 510–539, 2018, Publisher: Routledge, ISSN: 07421222 (ISSN). DOI: 10.1080/07421222.2018.1451956.

Paper session 3: AI for Language Analysis

Understanding Large Language Models through the Lens of Artificial Agency*

Maud van Lier[1]

Abstract— **This paper is motivated by Floridi's recent claim that Large Language Models like ChatGPT can be seen as 'intelligence-free' agents. Where I do not agree with Floridi that such systems are intelligence-free, my paper does question whether they can be called agents, and if so, what kind. I argue for the adoption of a more restricted understanding of agent in AI-research, one that comes closer in its meaning to how the term is used in the philosophies of mind, action, and agency. I propose such a more narrowing understanding of agent, suggesting that an agent can be seen as entity or system that things can be 'up to', that can act autonomously in a way that is best understood on the basis of Husserl's notion of indeterminate determinability.**

I. INTRODUCTION

For the past few months, many news items have been devoted to the recent developments in AI-research. Large Language Models (LLMs) like ChatGPT have stunned their users with their ability to produce texts that are almost indistinguishable from texts written by humans. Especially in the production of standard texts, it seems only a matter of time (if it is not the case so already) that these systems will outperform the majority of human writers in writing (standard) texts. At the same time, the mass use of LLMs has shown some of their flaws as well. Where the texts that LLMs produce might seem meaningful qua content *to us*, semantics plays no actual role in the word-generation processes of LLMs. Generated texts can thus contain a variety of mistakes that may not be obvious to us at first glance, making their uncritical use problematic.

In a recently published article, Floridi [1, pp. 4–5] states that "the implications of LLMs and the various AI systems that produce content of all kinds today will be enormous. (...) Some jobs will disappear, others are already emerging, and many will have to be reconsidered". What Floridi seems most interested in, though, is the many challenges that philosophers face in trying to understand the nature and role of this new kind of artificially created 'agents'. We have, according to Floridi [1, pp. 5–6], succeeded in creating a new form of agency, one where 'the ability to act' has been successfully decoupled "from the need to be intelligent, understand, reflect, consider or grasp anything". LLMs like ChatGPT are, then, what Floridi [1, p. 6] would call "intelligence-free agents". Even though I agree with Floridi that the societal integration of LLMs raises many new and challenging philosophical questions, I do not agree with him that this is because we have, in creating these systems, "liberated agency from intelligence" [1, p. 6].

In this paper, I will defend two claims. First, I argue that current LLMs are not 'intelligence-free' in the way that Floridi claims that they are — as not requiring cognitive processes to produce meaningful output — and that therefore we have not yet succeeded in liberating agency from intelligence (section II). Second, I argue that rather than focusing on whether or not LLMs are intelligence-free agents, the more interesting question is whether they can be called 'agents' in the first place. Where it is not unusual to refer to artificial systems as agents in computer science and robotics, these agents have had little to do with the kind of complex entities that are generally referred to as agents in the philosophies of mind, action, and agency. The recent successes of LLMs, however, suggest that this second, and more narrow understanding of agent might (soon) be applicable to LLMs (and other artificial systems) as well. Yet, to be able to determine what artificial systems could then be called such agents, and under which conditions, we need a proper understanding of what it means to be an agent in this more restricted sense. As I will show in section III, Floridi's agent account is too broad to be used for this purpose and I therefore propose a more restricted notion of agent that is able to capture what being an agent in this second sense entails.

II. COLLABORATIVE AGENTS

As mentioned in the introduction, Floridi [1, pp. 5–6] states that in creating AI-systems like LLMs, "we have decoupled the ability to act successfully from the need to be intelligent, understand, reflect, consider or grasp anything". In my understanding, Floridi means by this that LLMs are able to produce 'meaningful' (in the broadest sense of the word) content, without needing to understand why or how — or even *that* — this content is meaningful. LLMs can thus produce the same end-product (the text) as we can, without having to rely on the cognitive processes we use to produce this end-product. They can act without thinking — they are "intelligence-free agents" [1, p. 6].

In this section, I argue against this view by providing two objections to a labeling of LLMs as intelligence-free agents. First, I argue that current LLMs can only *act* in collaboration with (other) agents — humans — and that these other agents *do* make use of cognitive processes. This co-ability to act is thus not truly decoupled from intelligence. Second, and in line with this, I argue that the agent that can perform such a collaborative act is a *collaborative*

*Support by VolkswagenStiftung grant Az:97721 is gratefully acknowledged.

[1]M. van Lier is with the Philosophy Department, University of Konstanz, 78464 Konstanz, Germany maud.van-lier@uni-konstanz.de

agent that consists of a LLM and a human component. As long as a human is still supervising the act (or even part of this collaborative agent), the collaborative agent is not 'intelligence-free'. Taken together, these two objections then amount to a refutation of Floridi's claim that, at present, we have "liberated agency from intelligence" in creating LLMs [1, p. 6].

A. No Agency without Collaboration

In a recent book, Russo [2, p. 18] points out that "scientists use and interact with machines throughout the whole process of knowledge production". Russo emphasizes that this knowledge production is therefore a *co*-production that takes place in a techno-scientific *practice*, one in which both parties (scientist and machine) play a fundamental role. I now want to argue that the 'act' of producing meaningful texts by LLMs should be understood in a similar fashion: LLMs do not generate texts *by themselves*. In practice, they *co*-produce texts together with their user. I use the word 'production' rather than generation here, as it emphasizes the active roles of both the system *and* the user of this system in the production.

Floridi [1, p. 2] himself points out that in learning to use ChatGPT, one must, among other things, learn "how to use the right prompts (...), check the result, [and] know what to correct in the text produced by ChatGPT". Without a prompt, ChatGPT will not generate a text. Without the right prompt or critical feedback, ChatGPT will not generate *meaningful* texts. At the moment, then, ChatGPT, or any LLM for that matter, does thus not generate texts in isolation. Rather, producing (meaningful) texts in these cases is a *collaborative production*, one that necessarily involves both the (human) user and the system. Let us now return to Floridi's statement: "we have decoupled the ability to act successfully from the need to be intelligent, understand, reflect, consider or grasp anything" [1, p. 6] . Is this truly the case when LLMs generate texts? I would argue against this. At present, LLMs can only produce a meaningful text *together* with their user, where the user still needs to be able to understand and reflect on the consequences of a particular prompt and on what it means for a text to be correct or meaningful. Yes, the system itself can generate texts without understanding, but the entire *practice of text production* requires the prompts and feedback of intelligent users.

B. No Agency without Intelligence

Meaningful text production by LLM and user, when seen as an 'act', should thus be understood as a *collaborative* act. Only agents can act, so what kind of agent can perform this act?

In an attempt to close the 'responsibility gap'[1] that emerges when automated systems harm humans, Nyholm [5] shows the usefulness of the concept of *collaborative agency*. He argues that

> It does indeed make sense to attribute significant forms of agency to many current robotic technologies, such as automated cars or automated weapons systems. But this agency is typically best seen as a type of collaborative agency, where the other key partners of these collaborations are certain humans [5, p. 1203].

Nyholm calls all collaborators in a collaborative agency agents, but not all of them are independent or autonomous.[2] Collaborative agency is instead understood as a hierarchical model "where some agents within the collaborations [the automated systems] are under other agents' [the humans] supervision and authority" [5, p. 1203].

Given the earlier points made about the co-production of meaningful texts by LLM and user, I would say that this kind of text production is a paradigm example of an act performed by a collaborative agent. There is even a hierarchy where the user supervises the production by giving prompts and feedback. Calling current LLMs agents, I would say, is thus under the understanding of them being the *non-autonomous* and *non-independent* components of a collaborative agent. The acts of this collaborative agent are supervised by the human component, where this supervision still requires the human user to understand what makes a text meaningful, and what prompts would or would not work to create such a text.

Let us now return to Floridi's claim. Can we say that 'we have liberated agency from intelligence'? I would again at present argue against this. LLMs are by themselves neither independently nor autonomously able to produce meaningful texts. Rather, using Nyholm's concept, current LLMs are agents in the sense that they form a part of a collaborative agent. This agent, in its entirety, is not intelligence-free, because its human-component supervises and guides the act that is performed.

III. ARTIFICIAL AGENTS

In the previous section I pointed out that, given the current involvement of human users in the eventual *meaningful* output of LLMs, a labeling of these systems as 'intelligence-free' agents seems to misrepresent the actual practice of their use. This was not to say, however, that what these systems can do is not a big step forwards in the development of more autonomous and advanced AI-systems. Quite the contrary: current LLMs show more than ever the need for the right conceptual tools to capture what it is that they can do better than, or different from, other systems, and what (dis)similarities they do or will share with us humans. What makes them interesting, though, is not that they can perform tasks without 'intelligence', but rather that they can carry out very complex tasks in an almost completely independent

[1]The term 'responsibility gap' refers to the fact that "there is an increasing class of machine actions, where the traditional ways of responsibility ascription are not compatible with our sense of justice and the moral framework of society because nobody has enough control over the machine's actions to be able to assume the responsibility for them" [3, p. 177]. For a critical view on whether or not there are such gaps in the first place, see also [4].

[2]Nyholm [5, p. 1202] states that his paper "questions the tendency to view these types of agency [the ones attributed to types of robots] as instances of autonomous or independent agency".

manner. This makes them appear more and more agent-like to us.

As mentioned in the introduction, using the term agent to describe artificial systems in computer science and robotics is quite common, but this use has had little to do with how the term is generally understood in the philosophies of mind, action and agency. Agents are understood in these fields as entities that things can be 'up to', that can behave autonomously in some sense. The quality of the output of LLMs, as well as the relatively simple prompts that they need to generate very complex output, suggest that — at least in the near future — we might be able to develop systems that can properly be called agents in this philosophical sense.

However, for us to actually be able to make an agent/non-agent distinction between artificial systems, it seems that we need a notion of agent that is more restricted than Floridi's use of the term. This is because, in Floridi's account, every system that has any kind of transformative effect on its environment can be called an agent, meaning that every artificial system already counts as an agent.

In this section, I will first present Floridi's notion of agent and compare it to an understanding of agent that is loosely based on agent accounts in philosophy. I show that Floridi's concept of agent does not carry any explanatory power on its own and is therefore not the preferred choice for a study of artificial systems as agents. Instead, I propose a more restricted reading of agent, one that, *already on its own*, can provide us with insight into, and standards for, current and future AI-systems.

A. The Explanatory Power of Agency

In *On the morality of artificial agents*, Floridi and Sanders [6] defend the view that 'artificial agents' (understood as artificial systems) can be involved in moral situations both as moral patients and as moral agents. To make this argument, they first spell out what they mean by 'agent'. According to Floridi and Sanders, a human being (Jan) is an agent

> if Jan is a system, situated within and a part of an environment, which initiates a transformation, produces an effect or exerts power on it [the environment], as contrasted with a system that is (at least initially) acted on or responds to it, called the patient [6, p. 9].

As Floridi and Sanders [6, p. 9] state themselves, in this understanding of an agent, "there is no difference between Jan and an earthquake". The domain of systems that one can thus attribute agency to in this account is very large, since any system that causes some change in its environment counts as an agent.

There is a growing discussion about what kind of entities belong to the domain of agents, where the proposed entities range from humans as the paradigm case, to animals and organisms [7], [8], to groups [9] and artificial systems [10], [5]. A danger of extending this domain too far is that the term 'agent' as a categorizing concept loses much of its explanatory power. I would argue that this is the case in the account of Floridi and Sanders, since it has become difficult to meaningfully distinguish between for example an agent and a force, or an agent and a cause, when both humans and earthquakes are included in the domain of agents. After all, what a human and an earthquake have in common is that they are both natural systems that are subject to natural laws, whose continued existence affects some change in their respective environments. If this is the commonality that they share, would it then not be easier to call them both a cause or a force, or something similar? One could argue that both the human and the earthquake are the initiators or at least the initial loci for changes in their environment and that this is what makes them into agents rather than a cause or a force. But even here one could make the point that it is (fundamentally) implied in the meaning of a cause and a force that they initiate or affect something. Given the definition of agent that Floridi and Sanders provide then, it is not immediately clear what the added value is of referring to these systems as agents specifically.

This added value of the term agent is better captured in most agent accounts in the philosophies of mind, action, and agency.[3] In such accounts, an agent generally refers to an entity that can act: it can do things that are up to it, and in this sense it can act autonomously. This agent is then contrasted with an entity to which things merely happen, whose behavior — if it exhibits any — is non-autonomous in the sense that it never requires any active involvement of the entity itself. In this understanding of agent, the 'active involvement' (typically seen as goal directedness, intentionality, or control [11, p. 8]) allows one to meaningfully distinguish between an agent and a cause or a force: where both Jan and an earthquake cause changes in their environment, only Jan seems to have some sense of control over *how* he changes this environment.

In Floridi's understanding of agent, no such distinction can be made except by adding an adjective to the term agent that specifies what *kind* of agent we are talking about. Here one can think of examples like a *rational* agent, a *moral* one, or, in Floridi's latest article, an *intelligence-free* one. In his account, then, the explanatory power is shifted from the term agent itself to the adjectives that one uses in combination with the term. This is of course not a problem. The point that I want to make, however, is that this shift is unnecessary since the term agent, in a narrower understanding that comes closer to its use in the philosophies of mind, action, and agency, can carry meaning already on its own. And combining the term with the aforementioned (as well as other) adjectives will only enrich this meaning.

Before introducing my agent account, I first want to show in more detail why adopting a more narrow understanding of agent can be beneficial to AI-research. For now, I will assume that such an account requires at the minimum that certain things can be up to the agent — that it can act autonomously in some sense. Already with this provisional definition, we can (in principle) make a distinction between artificial systems that Floridi and Sanders cannot. In their

[3]See for an overview of the various positions in these fields [11] and [12]

account, both automated system and autonomous systems are agents as long as they initiate a change in their environment in some way. Any distinction between automated systems and autonomous ones can thus not be made clear by them by using the term agent alone.

I have said that this distinction can be made in principle, because we have gained nothing when we do not make clear what we mean by 'autonomous' behavior, especially when contrasted with 'automated' behavior. This distinction is not as clear-cut, because where mere automated systems cannot display autonomous behavior, autonomous systems do display automated behavior quite often. Just think of ourselves, the most paradigmatic agents out there. Much of what we do (and are doing) can be seen as automated behavior. When I walk while thinking for example, or when I am breathing. I could influence these doings, but do not necessarily have to. It is thus not that agents do *not exhibit any* automated behavior, but that they are capable of autonomous behavior too.[4] Where the difference between automated and autonomous behavior seems quite intuitive in living beings,[5] the distinction is less obvious in the case of (complex) artificial systems.

Let me try to explain what I mean. A recent advancement in fields like chemistry and materials science is the self-driving laboratory. Such a laboratory consist of an AI-system and a robotic platform, where the AI-system controls the robotic platform. The functioning of these laboratories is described by researchers as 'autonomous': the AI-system allows for the automation of both the design and execution of very specialized experiments within the limits of this robotic platform [14], [15]. Compared to systems that are merely automated, these laboratories are autonomous in that they can take over tasks that previously only the researcher could do, like the generation of hypotheses, and the design of, and control over, experiments that can test these hypotheses. Here it becomes interesting though, because where, compared to automated systems these self-driving laboratories seem to function autonomously, the question is whether we would continue to say so when we compare their functioning to our own way of doing things.

As said before, the term agent in philosophy generally refers to a system that things can be 'up to'. What does this mean in the case of the self-driving laboratory? The answer is not as clear. Self-driving laboratories function independently within a highly specialized context, where the number of experiments that can be executed is limited by, for example, the degrees of freedom of a robotic arm, or the overall technologies included in the platform. What is more, there is often an already pre-determined way to go about the design of experiments and the generation of hypotheses. Self-driving laboratories can take over routinized scientific practices. Where before only the experiment itself was automated, now the entire *practice* is automated, so both the technological and the human part, allowing for automation on a higher level. On the one hand, then, these systems function autonomously in the sense of executing a practice independently, while on the other hand they are limited in how they execute this practice given their training and the narrowness of the scope in which they function. Given these latter features, one could also make a case for a description of their functioning as one of high-level automation. The question then becomes whether being capable of the former kind of autonomy — independent functioning — is sufficient for being called an artificial agent in the narrow sense since it could be described as high-level automation too.

My aim here is not to speak in favor of either of the two options. What I want to point out, though, is that this is a fruitful discussion. One that is the result of adopting a more narrow understanding of agent, and then discussing whether a system could appropriately be called such an agent or not. It is a discussion that forced us to think about what we mean by automated and autonomous functioning in artificial systems, a domain of systems in which this distinction is less intuitively obvious than when talking about living systems for example. It suggested as well that functioning might be autonomous *up to a degree*. The question is now what we mean concretely when we say that things can be up to an agent. In the following, I will provide a more specific reading of this notion and reflect on how it can be used to think about LLMs and other artificial systems.

B. Autonomy as Indeterminate Determinability

Thus far, I have worked with an understanding of agent that, even though it is more narrow than Floridi's use of the term, is still rather abstract. At the beginning of section III-A, I defined agents as entities that things can be 'up to', that are autonomous because it is the agent and nothing else to which certain things are up. This notion of agent is derived from the account of Steward [8][6] and has served me well until now because it is broad enough to potentially include artificial agents,[7] but is not so broad that the term no longer carries any explanatory power of its own. Of course, by saying that things can be up to the agent, nothing much has been said yet. Where we might have an intuitive sense of what this means, we need a more workable definition to determine what artificial systems can or could count as agents — we need to know what the kind of 'things' are that can be up to agents, and how such things can be up to them.

Agents, according to Steward, are able to 'settle things' in the sense that "any exercise of agency is always such that it does not have to have happened" [8, p. 104]. She claims that (much of) the animal kingdom can already count as such agents. There are many parts of Steward's account that

[4]Wu [13, p. 203] states that "to understand agency, we must see that every process can have automatic and controlled features". Even though I do not necessarily see in Wu's account how one can make a clear distinction between the two features, I think that it is important to recognize that every agent can have more or less control, but minimally has to have some control.

[5]One could refer to things as 'autonomous behavior is the behavior one is aware and/or conscious of', or compare autonomous behavior with mere reflexes.

[6]"Agents are entities that things can be up to" [8, p. 25].

[7]Even though Steward is sceptical about the possibility of us being able to create artificial agents, her position does not reject it as an option [8, See for example p. 15 & footnote 40].

capture what we intuitively associate with agents, and that should therefore be included in an agent account. However, it also makes an objection clear that any such account faces, including my own: the question of whether there can even be something like agency in a world that appears to be mostly deterministic. Even though answering this question exceeds what is possible in this paper, Steward has made a nice attempt to do so, and I will shortly touch upon it here.

Steward's overall aim is to defend an Agent Incompatibilist position. Agent Incompatibilists hold that agency itself is incompatible with universal determinism. Where most Agent Incompatibilists have argued that *human* agency is incompatible with universal determinism, Steward argues that *animal agency* makes already trouble for universal determinism. Her argument is that the overwhelming number of examples of entities that things seem to be up to — that are capable *making themselves move* rather than move by themselves — make it very likely that agents exist and that therefore universal determinism is not true [8, see pp. 12-15]. Even though I find Steward's arguments convincing, they are based on intuitive and logical reasoning, and not on any conclusive empirical proof. I will therefore not make any claim about whether things that appear to be up to certain systems are truly 'up to' them. I will merely hold that if it appears like they do, then it makes sense to refer to these systems as agents.

So what does Steward mean when she says that things can be 'up to' agents? As stated before, she holds that things can be up to the agent in the sense that it is able to settle things. This does not mean however that the agent is free to do whatever it wants. Its ability to act is constrained both by its nature as well as by its environment. According to Steward, "it is utterly undeniable that all animal agency takes place within a framework which constrains, sometimes very tightly, what can be conceived of as a real option for that animal" [8, p. 104]. Within these constraints, though, agents have a lot of flexibility. Even though an animal has to eat within a certain time frame, for example, it can still decide when to eat, what route to take to get there, to eat slow or fast, to grasp its food this way and not that way, etc. The agent thus continuously settles *how* it does the things that it does. More advanced agents like us might have the ability to also settle more of *what* we do, but at the minimum for Steward an entity has to be able to settle how it does things for it to be an agent.

I think that there are two important things to take away from this account.[8] A first thing is that both the nature of the agent and its environment constrain what can be 'up to' it. Even though Steward focuses mostly on constraints that have to do with the particular embodiment of the agent and the natural laws that it and its environment are subject too, I think that a third factor that should be included is personal history. We evolve and learn over time, and each learning trajectory is different. The three factors that influence the kind of things that the agent can settle are thus the kind of agent that it is, its personal history, and its environment.

A second important aspect of agents is that they are *unpredictable* in a sense: even though the 'real options' available to them might be limited because of the above-mentioned factors, it is still the agent that settles what it will do in the end. Given that it is the agent that settles, we cannot exactly predict what it will do. This does not mean that their behavior is *random*, though. The options available to them depend on in part on the kind of agent they are, their personal history and the environment in which they move. So how can the autonomy of an agent be understood if its behavior is both predictable and unpredictable?

I think that Husserl's notion of *indeterminate determinability* [16, p. 283] can be quite illuminating here. Even though Husserl talks about this notion in the context of our Ego and what makes us a person, I find the notion helpful to spell out exactly what kind of unpredictability agents exhibit. Husserl uses the notion of style to indicate the kind of stereotypical behavior that we can expect from other people. Through our lived experiences, we develop certain styles of behavior that each present us with a number of options to behave in particular situations. However, since each of us has our own lived experience, we all have our own personal mixture of styles, we are each our own "individual kind" [16, p. 286].

The fact that we are our own individual kind makes our behavior unpredictable up to an extent:

> One can to a certain extent expect how a man will behave in a given case if one has correctly apperceived him in person, in his style. The expectation is generally not plain and clear; it has its apperceptive horizon of *indeterminate determinability* within an intentional framework that circumscribes it, and it concerns precisely one of the modes of behavior which corresponds to the style [16, p. 283, italics are mine].

The autonomy of an agent can thus be understood as indeterminate determinability: the agent will behave according to stable patterns, but can always diverge from them. These stable patterns depend in part on the agent's personal history, in part on the kind of system it is, and in part on the world that it moves in. The influence of each of these three factors on the action courses that are available to the agent, make each agent into its own 'individual kind'. Given that we all move through the same world and share certain characteristics with each other, we develop stable patterns (or styles of behavior) that resemble those of others. These patterns make it easier for other agents to predict what we will do, but since we are agents, we are also unpredictable in that in the end *the agent* settles what it will do.

In this more narrow understanding of agent, can we say that current LLMs can be seen as agents? I think that arguments can be made both for the affirmative, and the negative. Given that the probability calculations on which the text generation of LLMs is based contain stochastic elements, this

[8] Actually I would argue that there is also a third point, namely that both the agent as well as its environment should contain stochastic elements. Even when it cannot be proven beyond doubt that our world is (in part) indeterministic, we respond to it, and the entities in it, as if it is.

production could be described as indeterminate (stochastic elements) determinability (probability calculations). A first question that can be raised is whether what we are judging to be predictable is our own behavior or that of the LLM. Is it just learning to simulate us as well as possible, or is it learning to generate texts on its own? So, is it the LLM that is predictable or are we?

As a personal history, one could argue that the LLMs gain experience through the training on a particular text-corpus.[9] The question is of course whether this suffices for a 'personal' history, or whether we need something more, like interaction with other agents or interaction with a physical world. A question is as well what kind of system an LLM is. What role does the hardware form in its functioning? In what way does it provide or constrain the options of the LLM? As for the software, is the LLM always only able to choose from the same number of 'options' for the next word? Or does this depend on the context? And, does it always need to choose? It seems that agents, to have the choice in how they respond, need to be able as well to not act — to not generate a text. Can LLMs choose not to respond? Should this be necessary for them to count as agents? Can they settle things on their own without being prompted? And is the choice process (even though complicated) always the same for the LLM? Or will it depend on the interaction that is subjected to?

A final question is what counts as an environment for these LLMs. One could say that their environment consists of the prompts of their users and maybe even any texts that they have access too. If they are connected to the internet, for example, is then the whole of the internet their environment? And do we include the hardware in the environment? Why yes or no? Is it important to limit this environment? What then count as the agent? Only the software, software and hardware, or the environment as well?

These are relevant and important questions, all prompted by adopting a more narrow understanding of agent and using it to study current artificial systems. Given that already this still rather crude notion of agency as indeterminate determinability can raise and guide many interesting questions that we might have, I think it is important that we should make use of a more narrow understanding of the term agent in AI-research.

IV. CONCLUSION

The successes of LLMs like ChatGPT are very impressive and foreshadow great changes in how we do things in our society. At the same time they also foreshadow a change in our interaction with AI-systems that we are not yet conceptually ready for. Whether we would call such LLMs agents or not influences namely the way we interact with them. If the artificial system is exactly that — a system — then my behavior will, and should, be different towards it than when it is an agent. A tool can break or malfunction and its use should therefore be regulated and learned.

Where we use tools, we *inter*act with agents, and this interaction is shaped by the kind of agent that we interact with. Think for example of a cat, that cannot be held responsible for damaging the couch with its scratches, because we do not expect it to be able to reason why what it did is making us agitated. A twelve-year old child, though, can be held responsible for writing on the wall, because we expect it to be able to understand why this is not okay. Where current LLMs are very impressive, we are not yet sure if they are agents, and if they are, what kind of agents. To make this clearer, we need to develop the notion of 'artificial agent' further. This will not only protect the users by telling them what kind of interaction they can expect, but also provide us with some conceptual tools for Explainable AI and AI-ethics. The concept of 'artificial agent' can do a lot of work, if we give it the attention it deserves.

[9] I want to thank my reviewer for pointing this out.

REFERENCES

[1] Floridi, L. (2023). Ai as agency without intelligence: On chatgpt, large language models, and other generative models. *Philosophy & Technology*, 36(1):15.

[2] Russo, F. (2022). *Techno-scientific practices: an informational approach*. Rowman & Littlefield.

[3] Sparrow R. (2007). Killer robots. *Journal of Applied Philosophy*, 24(1):62–77.

[4] Tigard, D. W. (2021). There is no techno-responsibility gap. *Philosophy & Technology*, 34(3):589–607.

[5] Nyholm, S. (2018). Attributing agency to automated systems: Reflections on human–robot collaborations and responsibility-loci. *Science and engineering ethics*, 24(4):1201–1219.

[6] Floridi, L. and Sanders, J. W. (2004). On the morality of artificial agents. *Minds and machines*, 14:349–379.

[7] Burge, T. (2009). Primitive agency and natural norms. *Philosophy and Phenomenological Research*, 79(2):251–278.

[8] Steward, H. (2012). *A metaphysics for freedom*. Oxford University Press.

[9] List, C. and Pettit, P. (2011). *Group agency: The possibility, design, and status of corporate agents*. Oxford University Press.

[10] Gunkel, D. J. (2012). The machine question. *Critical perspectives on ai, robots, and ethics*, page 5.

[11] Ferrero, L. (2022). *The Routledge Handbook of Philosophy of Agency*. Routledge.

[12] Paul, S. (2020). *Philosophy of action: A contemporary introduction*. Routledge.

[13] Wu, W. (2022). Agency, consciousness and attention. In Ferrero, L., editor, *The Routledge Handbook of Philosophy of Agency*, chapter 18, pages 201–210. Routledge, Taylor & Francis Group, first edition.

[14] Seifrid, M., Pollice, R., Aguilar-Granda, A., Morgan Chan, Z., Hotta, K., Ser, C. T., Vestfrid, J., Wu, T. C., and Aspuru-Guzik, A. (2022). Autonomous chemical experiments: Challenges and perspectives on establishing a self-driving lab. *Accounts of Chemical Research*, 55(17):2454–2466.

[15] Abolhasani, M. and Kumacheva, E. (2023). The rise of self-driving labs in chemical and materials sciences. *Nature Synthesis*, pages 1–10.

[16] Husserl, E. (1989). *Ideas pertaining to a pure phenomenology and to a phenomenological philosophy: Second book studies in the phenomenology of constitution*, volume 3. Springer Science & Business Media.

Towards Better Product Quality: Identifying Legitimate Quality Issues through NLP & Machine Learning Techniques

Rakhshanda Jabeen[1,2], Morgan Ericsson[2] and Jonas Nordqvist[3]

***Abstract*—Manufacturers of high-end professional products are committed to delivering outstanding customer-quality experiences. They maintain databases of customer complaints and repair service jobs data to monitor product quality. Analyzing the text data from service jobs can help identify common problems, recurring issues, and patterns that impact customer satisfaction, and aid manufacturers in taking corrective actions to improve product design, manufacturing processes, and customer support services. However, distinguishing legitimate quality issues from a brief, domain-specific text in service jobs remains a challenge. This study aims to automate the classification of technical service repair job data into legitimate quality issues or non-issues to assist individuals in the quality field department in a large company. To achieve this goal, we developed a comprehensive pipeline based on natural language processing and machine learning techniques including raw text preprocessing, dealing with imbalance class distribution, feature extraction, and classification. In this study, We evaluate several feature extraction and machine learning classification methods and perform the Friedman test followed by Nemenyi post-hoc analysis to find the best-performing model. Our results show that the passive-aggressive classifier achieved the highest average accuracy of 94% and 89% average macro F1-score when trained on TF-IDF vectors.**

I. INTRODUCTION

With the rapid growth of unstructured data in electronic text formats, natural language processing (NLP) — a subfield of linguistics, computer science, and AI, has emerged as a vital field of research. NLP enables machines to understand and interpret human language. Companies are beginning to recognize the economic value of their text data repositories, including social media platforms and internal document collections, for informed decision-making [1].

The text classification task is one of the most essential tasks in NLP. It involves the automatic categorization of text documents into predefined classes based on their content using machine learning (ML) methods. The process generally includes several steps, including preprocessing (which involves tokenization, stopwords and noise removal, and lemmatization [2]), feature extraction (which involves converting natural language into numerical vectors for mathematical computation), and finally, modeling the data using an appropriate machine learning algorithm for classification. These techniques have a wide range of applications in various industries, such as healthcare, the Internet of Things (IoT), security, spam filtering, digital marketing, and sentiment analysis [3, 4].

This research aims to address the challenge faced by a multinational professional appliance manufacturing company, in the manual categorization of service repair data of machines to filter the legitimate quality issues in service jobs. Technical service agents are responsible for resolving customers' issues and providing a text description of the resolution. The quality field department then manually assesses and classifies the service jobs to determine if it is a genuine quality issue and requires attention at the production and design levels. However, with a growing volume of data in multiple languages, manual classification has become increasingly complex. Therefore, an automatic solution for the classification process is necessary to save time and resources while ensuring consistency and accuracy.

Wang et al. [5] proposed a medical triage system that uses NLP and ML methods to classify questions of patients and text related to their symptoms and to provide suggestions on which consulting room to choose. The system can potentially alleviate the burden on the hospital triage system by helping with disease diagnosis. Additionally, ML-based applications have been found to perform equally or better than individual clinicians, resulting in reduced time and resource requirements for the task [6].

Text classification techniques have also been used to detect spam emails due to the increase in the volume of emails. Spam filters can be implemented at various levels, such as client-level and email servers. Researchers have proposed a significant body of work to automatically and efficiently classify emails as spam or non-spam using NLP and ML methods. Such applications have proven to be effective in reducing the negative impact of spam emails, including wasted time and resources, financial losses, and phishing attacks [7, 8].

NLP and automatic text classification have numerous applications in the financial industry, including fraud detection, stock market predictions, investment recommendation, financial risk management, etc. [9]. Nair et al. [10] proposed a method that uses sentiment analysis of news headlines extracted from the Cryptopanic API to predict the price of Bitcoin. This system has the potential to help novice and professional traders make more profitable investment decisions and reduce the risks associated with cryptocurrency trade.

The successful application of text classification in various industries motivates us to propose an ML-based solution to automatically classify the industrial domain-specific text data effectively. The goal is to automate the classification process of technical service repair jobs data as either legitimate quality issues or not. This research work is focused on answering the following research questions:

1) Propose and implement a comprehensive preprocessing protocol that can effectively address the challenges asso-

[1]Electrolux Professional AB, Sweden
[2]Department of Computer Science and Media Technology, Linnaeus University, Växjö, Sweden
[3]Department of Mathematics, Linnaeus University, Växjö, Sweden

ciated with the dataset, particularly the presence of very brief, multilingual, and domain-specific text data with an imbalanced class distribution.
2) Determine the best-performing feature extraction method for the classification task and investigate its impact on the performance of the classifiers.
3) Evaluate the effectiveness of different machine learning classifiers in classifying service jobs as either genuine quality issues to be approved or non-quality issues to be rejected.

The organization of the paper is as follows. Section II provides an overview of related work, while SectionIII presents a detailed description of the problem. Section IV describes the methodology used to achieve the research objectives, including the description of the data set, text preprocessing techniques, feature extraction approaches, classification algorithms, and performance metrics used to evaluate the classification model. Sections V and VI present the results obtained and their discussion, respectively. Finally, Section VII concludes the work.

II. RELATED WORK

Numerous research studies have been conducted in NLP and ML, aiming to understand human language and address various real-world challenges. Among these efforts, text classification has been one of the primary focuses [11].

For data augmentation in text classification tasks, Wei et al. [12] proposed easy data augmentation (EDA) techniques, including synonym replacement, random insertion, random swap, and random deletion while preserving the original class labels. Their results showed improved classification accuracy on several benchmark datasets and reduced overfitting, especially when trained on smaller datasets. Fromme et al. [13] introduced ContextGen to address the challenge of low-resource domain-specific text classification tasks. They adapt the GPT-2 text generation model to generate domain-specific text samples and then assign labels to these generated samples using BERT to augment training input.

The study by Parmar et al. [14] compares the performance of five classifiers, including Support Vector Machines (SVM), multinomial Naive Bayes, Decision tree, Random forest, and k-nearest neighbors (k-NN), to categorize defects and issues reported in customer complaint messages in an industrial setting. They used the TF-IDF vectorizer for feature extraction and found that SVM achieved the highest accuracy of 63.02% among all classifiers.

Ohata et al. [15] proposed a modular pipeline for a technical support system to automatically categorize incoming customer issues and recommend appropriate solutions based on textual descriptions of the issues. One of the challenges associated with this study is the predominance of short messages and the presence of domain-specific technical terms in the dataset. The authors evaluated and compared various text representation techniques and ML classification methods and found that the Random Forest classifier achieved the highest accuracy of 72.7% and a weighted F1-score of 69.2%.

Hoffimann et al. [16] proposed an automatic classification approach to categorize the daily reports of drilling engineers into three classes aiming to reduce accidents, improve drilling companies' efficiency, and make informed decisions. The challenges present in the text corpus include technical symbols, abbreviations of technical terms, misspellings, and truncated sentences. They used skip-gram a variant of Word2Vec embeddings for feature extraction and three neural networks and found that LSTM performs best for the task with an average accuracy of 82.7%.

III. PROBLEM STATEMENT

At Electrolux Professional, the customer support centers collect information on customer complaints about machines installed in 17 countries in text form. The customer support center assigns the task of addressing these complaints to a technical service agent, who visits the site, resolves the issue, and records the service job in the database. The service agent provides a text description of the resolution, including details of any faulty components and defects in the machine, selected from a pre-defined set of codes. In the subsequent phase, the individuals in the quality field department manually evaluate the job to determine whether it constitutes a genuine quality issue that requires attention at the production and design levels. This evaluation involves categorizing the service jobs into two groups: approved as a quality issue or rejected as not a quality issue. The rejected calls are further categorized into different categories that include maintenance that should be done by the customer, installation fault, customer misuse of the product, and consumables that customers should replace.

As the company expands its business to various locations, it acquires more data related to service jobs. During the past five years, the quality field department has manually classified service jobs in various languages during the machine warranty period. In our preliminary analysis of historical data, we discovered that a significant number of rejected and approved service jobs share similar characteristics. Automating the classification of technical service job text can help save time and resources while also ensuring consistency and accuracy in the classification process. Additionally, an automated approach can quickly identify quality issues and prioritize them based on their severity and impact on customer satisfaction. Therefore, in this study, we propose an ML-based solution to automate the classification process for the quality field department.

One of the challenges associated with the dataset is the installation of machines in multiple countries and the use of local service agents to resolve the issues, resulting in various reporting styles and languages. The data set mainly comprises text generated by human operators and technicians, which contains instances of misspellings, abbreviations, and inconsistencies in the representation of faulty component codes for similar tasks. We have implemented a detailed preprocessing protocol to address these challenges, which is outlined in the following section.

IV. METHODOLOGY

In the following sections, we will elaborate on the schematic workflow of the proposed methodology, as shown in 1, which outlines the classification process for service jobs. This methodology is composed of several phases, including preprocessing, addressing the issue of imbalanced class distribution, feature extraction, and classification. Finally, the performance metrics are evaluated and the analysis of the results is presented.

A. Dataset

The dataset used in this study is derived from the Electrolux Professionals service records database related to the laundry machines within the warranty period of machines. The service jobs have been manually reviewed and approved or rejected by the field quality team for the last five years. The dataset included customer complaints, technical comments in 17 languages, and other relevant metadata for each service job. Around 84% of these jobs are approved as quality calls when manually annotated, while nearly 16% of jobs are rejected, indicating a considerable imbalance between the two classes.

B. Preprocessing

The preprocessing phase is a fundamental step in Natural Language Processing (NLP) and machine learning classification, as it helps to comprehend the desired outcome. This study deals with unstructured, brief multilingual text data that lacks sufficient information for the classification task. The dataset includes categorical features like faulty component codes, defect codes, and the cost of replaced components, along with textual features technician reports, and customer complaints. Our analysis revealed that adding further categorical features and associated metadata could significantly enhance the predictive strength of the model. Therefore, we mapped faulty component codes to their corresponding names, main groups, and subgroups, considering the possibility of a single name referring to either an electrical or mechanical component. We then combined these informative categorical features with our textual features, such as customer complaints and technical comments, to create a single document for each service job. This approach aimed to improve the overall predictive strength of the model by utilizing both categorical and textual information comprehensively.

This data presents multiple challenges, including multilingual text, missing or incomplete sentences, incorrect spellings, and varying terminology and abbreviations used by different technicians in the reports. In addition, the encoding of special characters from non-English languages in older reports is flawed and has been replaced with symbols such as #, {, $, etc. To address these limitations, we conducted a manual search to identify and correct common misspellings and aimed to establish a uniform language representation of text in all languages. To achieve this, we created a dictionary that maps domain-specific technical terms and abbreviations from all languages to English and used the Google Translation API to translate the text data into English.

In summary, the preprocessing techniques involve common misspellings corrections, translation, conversion to lowercase, tokenization, lemmatization, and removal of stopwords, punctuation, and extra white spaces. Afterward, we split the data into two sets with a ratio of 80% for training and 20% for testing the model.

C. Dealing with Imbalanced Class Distributions

The dataset represents an imbalanced class distribution, with the class "approved" being overrepresented. Such an imbalance in the classes can hinder the generalization of classification models, as they might be biased toward the overrepresented class. Following the strategy of Wei et al. [12], we used random synonym replacement to augment the minority class. This technique involves randomly selecting n nouns and verbs from a given document and replacing them with their corresponding synonyms using a certain probability. Specifically, we randomly selected a subset of 60% training inputs from the minority class in the training dataset and applied the synonym replacement from WordNet [17] with a probability of 0.5 on all documents and then added these instances in the training set.

D. Feature Extraction

When dealing with NLP applications, finding an appropriate numerical representation of text data is essential to make it mathematically computable for a machine learning algorithm. There has been extensive research on various feature extraction methods for numerical representation of text [18, 19, 20, 21, 22]. Different feature extraction techniques highlight different features of the data and produce varying outcomes. Therefore, selecting a suitable representation of the text data significantly affects the text classification experiments. In this study, we have used four different feature extraction techniques to numerically vectorize the tokens of service jobs as follows:

1) Term frequency-inverse document frequency (TF-IDF) [23] is a conventional term-weighing technique to extract features from text data. TF-IDF captures the importance of a term in a document by assigning a high weight to terms that appear frequently in the document, but not so often in other documents.
2) Word2Vec [20] is a neural network-based predictive word embedding technique that can be further divided into two variants; the continuous bag of words (CBOW) and skip-gram. We utilized the skip-gram method that learns word vectors by training the network to predict the context of a word within a fixed window, given the word. In our experiments, we trained the Word2Vec model on our training data for 20 epochs, setting the context window equal to 5. Each word in the corpus was embedded into a 300-dimensional vector, and service jobs were represented as the average of their respective word vectors.
3) Doc2Vec [21] is a predictive document embedding technique that extends the concept of Word2Vec to generate embeddings for documents. The embeddings

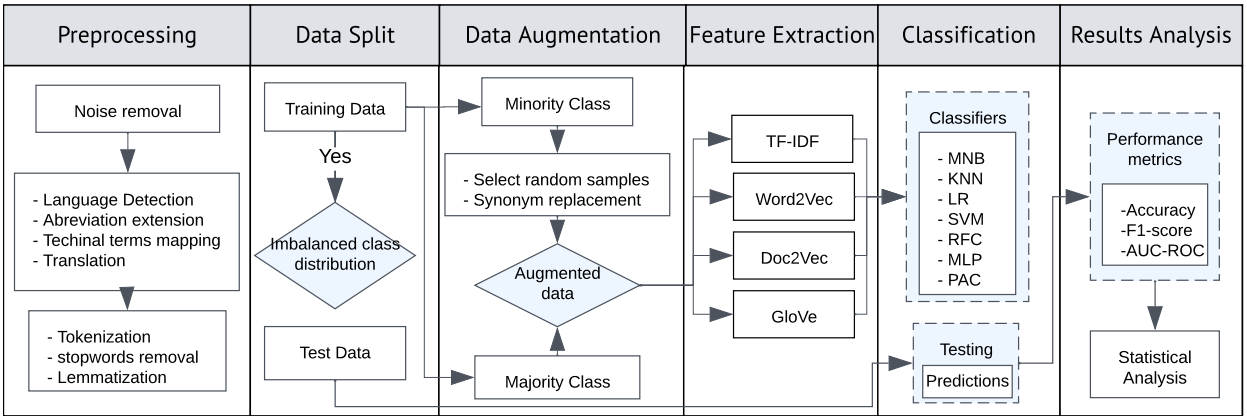

Fig. 1: The schematic workflow of the proposed methodology. First, the manually tagged historical data is preprocessed using various NLP techniques. Next, a data augmentation technique is applied to the minority class of the training data. Then feature extraction models are trained on training data and features are fed into one of the six classification models and then tested on the test data. Finally, a result analysis is presented to evaluate the performance of the model.

can be learned using either paragraph vector distributed memory (PV-DM) or paragraph vector distributed bag of words (PV-DBOW) model. In our work, we used DBOW to learn the document vectors from our training data and embed each service job into a 300-dimensional vector. The model was trained for 20 epochs with a context window of size 5.

4) GloVe (Global vectors for word representation) [22] is a count-based word embedding technique that leverages global word-to-word occurrence counts and statistical information to learn word vectors. Our study used a pre-trained GloVe word embedding with 300 dimensions. Similar to the Word2Vec approach, we obtained a vector representation of service jobs by taking the average of the word vectors for each word in a document, resulting in vectors with a dimension of 300.

E. Classification

Given the labeled service jobs data, we utilized supervised machine learning techniques for classification. This involved training a model on the labeled dataset and using it to make predictions on new data. In this study, we evaluated the performance of traditional and state-of-the-art machine learning models that are widely used in text classification, drawing inspiration from prior research that has demonstrated their effectiveness in various applications, such as sentiment analysis, spam filtering, and document categorization [24, 25, 26]. The classification models evaluated in this research are as follows:

1) Naive Bayes (NB) [27], a simple probabilistic algorithm that is computationally efficient and works well with high-dimensional data. However, NB assumes that predictor features are independent and uncorrelated, which may not hold in several real-world scenarios.
2) k-Nearest Neighbor (k-NN) [28], a non-parametric algorithm that classifies new instances based on their similarity to the k nearest neighbors in the training set. The algorithm makes no assumptions about the data distribution and can handle non-linear boundaries. however, it may be computationally inefficient and sensitive to the choice of hyperparameter k.
3) Logistic Regression (LR) [29], a statistical algorithm that utilizes the logistic function to predict the probability of each observation belonging to a certain class. LR is easier to implement and interpret and can handle large datasets. However, it assumes that data features are independent of each other and follow a linear relationship with the response variable.
4) Support Vector Machines (SVM) [30] is a non-parametric algorithm that finds a hyperplane that best separates the classes. This hyperplane is called the decision boundary. SVM can learn linear and non-linear boundaries between classes and is effective in high-dimensional data. However, it requires a careful selection of hyperparameters and takes a long training time on large datasets. Moreover, it is sensitive to the choice of the kernel function.
5) Random Forest (RF) [31] is an ensemble learning algorithm that combines multiple decision trees and its output is the mean prediction of individual trees. The RF algorithm can handle both linear and complex relationships and explicitly performs feature selection. However, RF cannot be easily interpreted and can be computationally expensive for large datasets.
6) Multi-Layer Perceptron (MLP) [32] is a feedforward neural network architecture that uses multiple layers of neurons (at least three layers including an input layer, a hidden layer of neurons, and an output layer). MLP can capture complex, non-linear patterns and relationships between data and works well on large datasets. However, MLPs are fully connected neural networks, resulting in too many hyperparameters that require careful tuning. In our experiments, we used one or two hidden layers.

7) Passive-Aggressive (PA) [33] is an online learning linear algorithm that learns incrementally and updates its weights based on a passive-aggressive strategy. The algorithm makes predictions based on the current weights when it receives a new input. If the prediction is correct, then the model acts passively, and the weights remain unchanged. If the prediction is incorrect, the algorithm acts aggressively and updates the weights to minimize the loss by adding a regularization parameter (C) that penalizes the weight vector. PA is known for its ability to adapt quickly to dynamic changes in the data. However, it cannot capture complex, nonlinear decision boundaries between classes and is sensitive to the choice of the regularization parameter.

The classification process is divided into two phases training and testing. In the training phase, 80% of the data is used to train the classification model. For this purpose, we created a pipeline that involves augmenting the minority class data to make it comparable with the majority class as mentioned in IV-C, followed by feature extraction and then classification using the classification algorithm to classify the service jobs. We optimized the model's performance by tuning the hyperparameter using an exhaustive search with 5-fold cross-validation (CV) over the hyperparameter configurations listed in Table I. This search was conducted for all combinations of feature extraction techniques and classification algorithms. The resulting optimal hyperparameters were then used to train the models. In the testing phase, the remaining 20% labeled data is vectorized using the feature extraction approach of the pipeline and then fed into the model for classification, and performance metrics are computed.

To evaluate the classification models, we used a 10-fold stratified CV method that divides the data into ten subsets of roughly equal size, with each subset imitating the class distribution in the entire dataset. We then trained and tested the model 10 times, using each subset once as the testing subset and the remaining subsets as the training subsets.

F. Evaluation Metrics & Statistical Analysis

When dealing with imbalanced class distribution in data, the accuracy score for evaluating the performance of models can be misleading. Therefore, it is necessary to select a metric that offers a more comprehensive assessment of the model and is insensitive to changes in data distribution. In this study, we have evaluated the predictive performance of classification models using the macro F1-score and the area under the Receiver Operating Characteristic (ROC) curve (AUC-ROC) scores.

The F1-score is the harmonic mean of precision and recall, where precision is the proportion of positive predictions made by the model and recall is the proportion of true positive cases correctly classified by the model. The macro F1-score is then determined by calculating the arithmetic mean of the F1-scores for all classes.

The ROC curve is a graphical representation of a classification model's performance obtained by plotting the true positive rate (TPR) on the y-axis against the false positive rate (FPR) on the x-axis. TPR is the same as recall, while FPR is the proportion of actual negative cases incorrectly classifies as positive by the model. The AUC-ROC score ranges from 0 to 1, where higher values indicate better performance.

To verify if there is a significant difference in the model's performance, we conducted Friedman test [34] at a significance level of 0.05 on the ranks of metric scores obtained from 10 CV runs of all classifiers. If the resulting p-value of the test statistic is less than 0.05, we reject the null hypothesis that all classifiers perform equally in favor of the alternative hypothesis that at least one classifier performs differently from the others. To identify which classifiers differ significantly, we performed the Nemenyi post-hoc test at a significance level of 0.05 and computed the critical difference (CD) based on the average ranks of all CV scores. If the mean ranks differ by at least CD, the two classifiers are considered significantly different.

G. Experimental Setup

In this study, we conducted 28 experiments exploring all possible combinations of feature extraction techniques and classification models described in the previous sections. The implementation was carried out in Python 3.10.6, with the aid of general machine learning and NLP tools such as scikit-learn [35], Gensim [36], spaCy [37], and NLTK [38].

TABLE I: Hyperparameters optimization setup for all classifiers

Classifier	General Setup	Hyperparameters	Grid setup
NB	multinomial	–	–
k-NN	–	no. of neighbors	$[3, 5, 7, 9, 11, 13, 15]$
LR	penalty: l_2	C	logspace$(-3, 3, 10)$
SVM	Kernel: rbf	C γ	$[0.1, 1, 5, 10, 100]$ $[0.001, 0.01, 0.1, 1]$
MLP	activation: $relu$ solver: $adam$	hidden layers	$[(10,), (50,), (100,), (10, 10), (50, 50), (100, 100)]$
RF	criterion: $gini$	no. of trees maximum depth	$[10, 50, 100, 200, 500]$ $[\text{None}, 10, 50, 100]$
PA	loss: $hinge$	C	$[0.001, 0.01, 0.1, 1, 10, 100, 1000]$

V. RESULTS

In this section, we present the results obtained from our experiments as described in IV. Table II reports the average scores and 95% confidence intervals (CIs) computed over 10 runs for each combination of the feature extraction method and classifier. The width of the CI reflects the level of uncertainty of the estimate, and a narrower CI indicates a more precise estimate. To obtain a reliable performance evaluation, it is desirable to have small CIs.

Based on the results, it is evident that the TF-IDF feature extraction approach yields the highest accuracy, macro F1, and AUC-ROC scores among all evaluated classifiers. Additionally, when combined with TF-IDF, LR, SVM, RF, MLP, and PA classifiers outperform NB and k-NN classifiers.

TABLE II: Accuracy, macro F1, and AUC-ROC scores for all combinations of feature extraction method and classifier

Feature Extraction	Classifier	Accuracy (%)	F1-Score (%)	AUC-ROC (%)
Doc2Vec	NB	88 ± 0.4	78 ± 0.6	77 ± 0.7
	k-NN	91 ± 0.9	83 ± 1.6	83 ± 1.7
	LR	90 ± 0.5	82 ± 1.4	87 ± 4.3
	SVM	92 ± 0.5	85 ± 0.9	86 ± 3.7
	RF	92 ± 0.5	85 ± 0.9	86 ± 3.2
	MLP	90 ± 1.3	83 ± 1.8	82 ± 3.0
	PA	90 ± 0.7	83 ± 1.5	83 ± 1.4
GloVe	NB	87 ± 0.4	70 ± 1.0	81 ± 1.5
	k-NN	90 ± 1.1	83 ± 1.9	82 ± 1.9
	LR	90 ± 1.1	83 ± 1.8	82 ± 2.1
	SVM	93 ± 0.8	85 ± 1.8	89 ± 1.5
	RF	92 ± 0.7	82 ± 1.3	90 ± 1.2
	MLP	92 ± 0.4	85 ± 0.9	86 ± 0.9
	PA	90 ± 0.6	83 ± 0.9	82 ± 1.1
Word2Vec	NB	88 ± 0.9	78 ± 1.9	78 ± 1.5
	k-NN	92 ± 0.2	86 ± 0.6	86 ± 0.2
	LR	92 ± 0.4	85 ± 1.0	85 ± 0.6
	SVM	93 ± 0.9	87 ± 1.8	90 ± 1.4
	RF	93 ± 0.6	86 ± 1.3	92 ± 1.8
	MLP	93 ± 1.0	86 ± 0.7	89 ± 2.1
	PA	92 ± 0.7	86 ± 1.3	85 ± 1.1
TF-IDF	NB	90 ± 0.8	83 ± 1.3	81 ± 1.5
	k-NN	91 ± 0.8	85 ± 1.4	85 ± 1.4
	LR	94 ± 0.9	88 ± 1.6	90 ± 1.8
	SVM	94 ± 0.7	88 ± 1.4	91 ± 1.2
	RF	93 ± 0.8	87 ± 1.5	90 ± 1.6
	MLP	94 ± 0.5	88 ± 0.8	90 ± 1.1
	PA	94 ± 0.6	89 ± 1.0	91 ± 1.3

A. Statistical Analysis

The experimental results presented in Table II demonstrate that a direct comparison of classifiers using the best feature extraction technique, TF-IDF, can be misleading, as there is minimal difference in accuracy, macro F1-scores, and AUC-ROC scores of the models. Therefore, we applied the Friedman test on the ranks of 10 CV runs of all classifiers when combined with TF-IDF. The results of the Friedman test are shown in Table III. All p-values are less than the significance level of 0.05, leading us to reject the null hypothesis and conclude that at least one classifier performs differently from the others.

TABLE III: Results of the Friedman Test on CV scores of all classifiers in combination with TF-IDF vectorization

Score	Friedman's statistics	p-value
Accuracy	34.66	10^{-6}
Macro F1	31.02	10^{-5}
AUC-ROC	36.43	10^{-6}

According to the Nemenyi post-hoc test, the CD is 3.404. The pairwise comparison of the average ranks of all classifiers, using TF-IDF vectorization, is presented in Figure 2. Groups of classifiers that are not significantly different (at the significance level of 0.05) are connected with a horizontal line, and the length of the line between the two classifiers indicates the difference between them. Furthermore, the lower the rank, the better the classification model in terms of the corresponding performance measure.

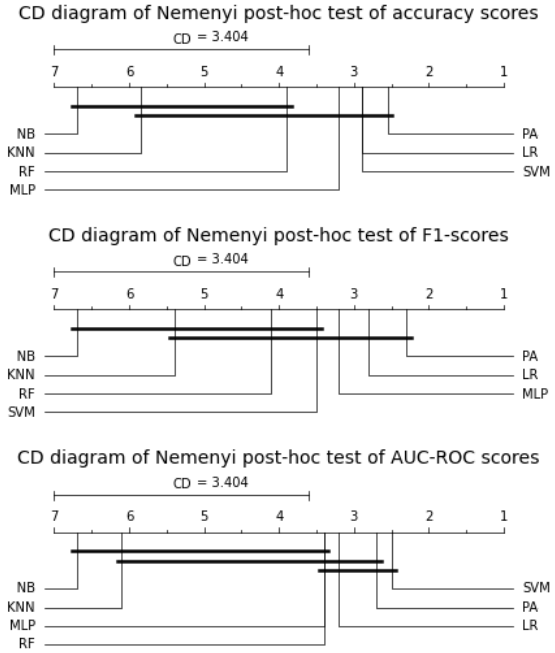

Fig. 2: Nemenyi test results on pairwise comparison on mean ranks of accuracy, F1 and AUC-ROC scores of all classifiers

VI. DISCUSSION

In this section, we will provide a conclusive analysis of the methodologies and results used to achieve the goals of this research. Finally, We will discuss the limitations of the methodology and datasets used.

A. Comparison of Feature Extraction Methods

The results in Table II show that, regardless of the paired classification model, Doc2Vec embeddings exhibited weaker predictive strength in terms of accuracy, F1-score, and AUC-ROC score than the other feature extraction techniques. The only exception was the NB classifier, which performed better

when trained on Doc2Vec features than GloVe. SVM and RF classifiers achieved the highest accuracy and F1-scores when trained on Doc2Vec embeddings, while LR classifiers achieved the highest AUC-ROC score. One possible reason for this could be that Word2Vec and Doc2Vec are unsupervised feature extraction techniques that require a large corpus of text to learn a meaningful representation of the text, which is not available in our small, domain-specific corpus.

Furthermore, the GloVe feature extraction method performed slightly better than Doc2Vec, achieving the best accuracy score of 93% using the SVM classifier. The highest F1-score of 85% was achieved by SVM and MLP, closely followed by the k -NN, LR, and PA classifiers. However, Word2Vec performed better than GloVe and Doc2Vec, achieving the highest accuracy score of 93% using SVM, RF, and MLP classifiers, followed closely by k-NN, LR, and PA classifiers with an accuracy score of 92%.

The TF-IDF feature extraction approach achieves the highest accuracy, F1-score, and AUC-ROC score across all evaluated classifiers. MLP, PA, and SVM classifiers achieved the highest accuracy score of 94% when using the TF-IDF feature extraction method. Furthermore, the PA classifier achieved the best macro F1-score of 89%, closely followed by the MLP, SVM, and LR classifiers. This could be attributed to the fact that the dataset contains domain-specific terms and abbreviations that are less frequent in general language usage. Furthermore, our documents have a more limited vocabulary, with a greater emphasis on specialized terms and jargon, causing TF-IDF to perform better as it assigns more weight to the terms that are specific to a certain domain.

In conclusion, the TF-IDF feature extraction method is the most effective method, irrespective of the choice of classifier in this task. However, Word2Vec and GloVe can also be used as effective alternatives to TF-IDF, particularly when using MLP and SVM classifiers.

B. Comparison of Classifiers

Based on the Nemenyi post-hoc analysis of the results of all classifiers when combined with the TF-IDF feature extraction approach (see figure 2), it can be inferred that PA, LR, SVM, and MLP achieved significantly higher accuracy scores than NB. Additionally, NB, k-NN, and RF classifiers did not show any significant differences in their accuracy scores. The Nemenyi post-hoc test on F1 scores showed that PA performed significantly better than NB, but no significant difference was found between PA, LR, MLP, and SVM in terms of F1-scores. Moreover, the Nemenyi post-hoc test on AUC-ROC scores revealed that SVM performed significantly better than NB and k-NN. At the same time, no significant differences were found between the SVM, PA, RF, and MLP classifiers.

The results in Table II suggest that the LR and SVM classifiers exhibit greater consistency across all feature extraction methods. However, the PA classifier achieves the highest accuracy and F1-score among all classifiers when trained on TF-IDF vectors. It is important to note that these findings may not generalize to all datasets and classification tasks, since the performance of different classifiers can depend on the unique characteristics of the data.

C. Potential Applications

Electrolux Professional's quality field department aims to consistently improve the quality of their machines by monitoring the KPI Service Call Rate (SCR). This study contributes to more accurate SCR calculations by identifying legitimate quality issues in service jobs, resulting in time and resource savings, as well as ensuring consistency and accuracy in the classification process

D. Limitations

During our experiments, we discovered several limitations that hinder the learning process of a machine learning model. Firstly, we observed inconsistencies in the labeling of service jobs, which could be attributed to the manual reading and labeling of the dataset by different quality team members. Secondly, customer complaints do not always provide sufficient information to accurately diagnose the machine issue. Furthermore, some technicians provide an insufficient job description, leading to a very short text and a lack of understanding of the problem's nature.

Another significant limitation that we encountered is the insufficient number of labeled data to train the model. The dataset has a skewed distribution, with the minority class divided into several subgroups, making it challenging to accurately classify. However, this limitation can be overcome by generating model predictions and manually reviewing instances with low confidence in the predictions. In summary, our experiments revealed that inconsistent labeling, insufficient information in customer complaints and technician descriptions, and a scarcity of labeled data are the primary factors that need to be addressed to improve the model's performance.

In addition to the techniques explored in this research, another potential solution for this problem could be found in transfer learning. Transfer learning involves training a model on a large, general dataset and then fine-tuning it on a smaller, more specific dataset. This approach is useful in real-world scenarios where we have limited labeled data, as it allows the model to leverage knowledge learned from a larger, more diverse dataset.

VII. CONCLUSION

The study aimed to automate the classification of technical service repair job data into legitimate quality issues or non-issues. In this concluding section, we present a comprehensive analysis of the answers to the research questions detailed in Section I.

To effectively address the challenges associated with the dataset, we proposed and implemented a comprehensive preprocessing protocol. One of the challenges was the predominance of very brief, multilingual, and domain-specific text data. To address this, we manually created a dictionary of technical terms and jargon used in the technicians' reports. We then used the Google Translation API in our preprocessing

pipeline to translate the data into English. In addition, to tackle the imbalanced class distribution problem, we implemented a data augmentation technique that randomly replaced some terms in a document with their synonyms at a certain probability to make the minority class comparable to the majority class. This preprocessing protocol has been effective in improving the overall accuracy of the classifiers.

We evaluated several feature extraction methods, including TF-IDF, Word2Vec, Doc2Vev, and GloVe, and found that TF-IDF outperformed the other methods for this classification task across all evaluated classifiers. TF-IDF's superior performance can be attributed to its ability to weigh technical terms and jargon effectively in domain-specific text. However, it is important to note that TF-IDF cannot capture the syntactic and semantic relationships between words and may suffer from sparsity when dealing with very short text documents.

We used seven machine learning algorithms in our study and found that the passive-aggressive classifier achieved the highest accuracy of 94% and an F1-score of 89% when trained on TF-IDF vectors. Therefore, we incorporated the PA classifier into our proposed framework, knowing that it can quickly adapt to rapid changes in the data and learns incrementally. However, PA assumes that classes are linearly separable and may not perform as well when dealing with complex data relationships.

In conclusion, the ML-based solution proposed in this study can effectively automate the classification of technical service job data, improve the efficiency of the quality field department, and save time. The techniques used in this approach can be extended to similar problems in other industries. Future work will focus on developing a solution for technicians and the customer care department to identify possible problem resolutions before visiting the site using only customer complaint data and historical records.

REFERENCES

[1] A. Ittoo, A. van den Bosch *et al.*, "Text analytics in industry: Challenges, desiderata and trends," *Computers in Industry*, vol. 78, pp. 96–107, 2016.

[2] M. Anandarajan, C. Hill, T. Nolan, M. Anandarajan, C. Hill, and T. Nolan, "Text preprocessing," *Practical text analytics: Maximizing the value of text data*, pp. 45–59, 2019.

[3] M. K. Dalal and M. A. Zaveri, "Automatic text classification: a technical review," *International Journal of Computer Applications*, vol. 28, no. 2, pp. 37–40, 2011.

[4] M. O. Aftab, U. Ahmad, S. Khalid, A. Saud, A. Hassan, and M. S. Farooq, "Sentiment analysis of customer for ecommerce by applying ai," in *2021 International Conference on Innovative Computing (ICIC)*, 2021, pp. 1–7.

[5] X. Wang, M. Tao, R. Wang, and L. Zhang, "Reduce the medical burden: An automatic medical triage system using text classification bert based on transformer structure," in *2021 2nd International Conference on Big Data & Artificial Intelligence & Software Engineering (ICBASE)*. IEEE, 2021, pp. 679–685.

[6] S. Delshad, V. S. Dontaraju, and V. Chengat, "Artificial intelligence-based application provides accurate medical triage advice when compared to consensus decisions of healthcare providers," *Cureus*, vol. 13, no. 8, 2021.

[7] S. Magdy, Y. Abouelseoud, and M. Mikhail, "Efficient spam and phishing emails filtering based on deep learning," *Computer Networks*, vol. 206, p. 108826, 2022.

[8] T. Peng, I. Harris, and Y. Sawa, "Detecting phishing attacks using natural language processing and machine learning," in *2018 ieee 12th international conference on semantic computing (icsc)*. IEEE, 2018, pp. 300–301.

[9] A. Gupta, V. Dengre, H. A. Kheruwala, and M. Shah, "Comprehensive review of text-mining applications in finance," *Financial Innovation*, vol. 6, pp. 1–25, 2020.

[10] K. Nair, A. Pawle, A. Trisal, and S. Krishnan, "Bitcoin price prediction using sentimental analysis - a comparative study of neural network model for price prediction," in *2022 2nd Asian Conference on Innovation in Technology (ASIANCON)*, 2022, pp. 1–4.

[11] Q. Li, H. Peng, J. Li, C. Xia, R. Yang, L. Sun, P. S. Yu, and L. He, "A survey on text classification: From shallow to deep learning," *arXiv preprint arXiv:2008.00364*, 2020.

[12] J. Wei and K. Zou, "Eda: Easy data augmentation techniques for boosting performance on text classification tasks," *arXiv preprint arXiv:1901.11196*, 2019.

[13] L. Fromme, J. Bogojeska, and J. Kuhn, "Contextgen: Targeted data generation for low resource domain specific text classification," in *International Conference on Artificial Intelligence and Statistics*. PMLR, 2022, pp. 3016–3027.

[14] P. S. Parmar, P. Biju, M. Shankar, and N. Kadiresan, "Multiclass text classification and analytics for improving customer support response through different classifiers," in *2018 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*. IEEE, 2018, pp. 538–542.

[15] E. F. Ohata, C. L. C. Mattos, S. L. Gomes, E. D. S. Rebouças, and P. A. L. Rego, "A text classification methodology to assist a large technical support system," *IEEE Access*, vol. 10, pp. 108 413–108 421, 2022.

[16] J. Hoffimann, Y. Mao, A. Wesley, and A. Taylor, "Sequence mining and pattern analysis in drilling reports with deep natural language processing," in *SPE Annual Technical Conference and Exhibition*. OnePetro, 2018.

[17] G. Miller, "" wordnet: An on-line lexical database, &quot," *International Journal of Lexicography*, vol. 3, no. 4, pp. 235–244, 1990.

[18] V. John, "A survey of neural network techniques for feature extraction from text," *arXiv preprint arXiv:1704.08531*, 2017.

[19] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint*

arXiv:1810.04805, 2018.

[20] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," *Advances in neural information processing systems*, vol. 26, 2013.

[21] Q. Le and T. Mikolov, "Distributed representations of sentences and documents," in *International conference on machine learning*. PMLR, 2014, pp. 1188–1196.

[22] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 2014, pp. 1532–1543.

[23] J. Ramos *et al.*, "Using tf-idf to determine word relevance in document queries," in *Proceedings of the first instructional conference on machine learning*, vol. 242, no. 1. Citeseer, 2003, pp. 29–48.

[24] R. Li, "A review of machine learning algorithms for text classification," *Cyber Security*, p. 226, 2022.

[25] X. Deng, Y. Li, J. Weng, and J. Zhang, "Feature selection for text classification: A review," *Multimedia Tools and Applications*, vol. 78, pp. 3797–3816, 2019.

[26] K. Nagashri and J. Sangeetha, "Fake news detection using passive-aggressive classifier and other machine learning algorithms," in *Advances in Computing and Network Communications: Proceedings of CoCoNet 2020, Volume 2*. Springer, 2021, pp. 221–233.

[27] I. Rish *et al.*, "An empirical study of the naive bayes classifier," in *IJCAI 2001 workshop on empirical methods in artificial intelligence*, vol. 3, no. 22, 2001, pp. 41–46.

[28] A. Mucherino, P. J. Papajorgji, P. M. Pardalos, A. Mucherino, P. J. Papajorgji, and P. M. Pardalos, "K-nearest neighbor classification," *Data mining in agriculture*, pp. 83–106, 2009.

[29] D. W. Hosmer Jr, S. Lemeshow, and R. X. Sturdivant, *Applied logistic regression*. John Wiley & Sons, 2013.

[30] M. Awad, R. Khanna, M. Awad, and R. Khanna, "Support vector machines for classification," *Efficient Learning Machines: Theories, Concepts, and Applications for Engineers and System Designers*, pp. 39–66, 2015.

[31] L. Breiman, "Random forests," *Machine learning*, vol. 45, pp. 5–32, 2001.

[32] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge, MA: MIT Press, 2016. [Online]. Available: https://www.deeplearningbook.org/

[33] K. Crammer, O. Dekel, J. Keshet, S. Shalev-Shwartz, and Y. Singer, "Online passive aggressive algorithms," 2006.

[34] K. Stąpor, "Evaluating and comparing classifiers: Review, some recommendations and limitations," in *Proceedings of the 10th International Conference on Computer Recognition Systems CORES 2017 10*. Springer, 2018, pp. 12–21.

[35] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg *et al.*, "Scikit-learn: Machine learning in python," *the Journal of machine Learning research*, vol. 12, pp. 2825–2830, 2011.

[36] R. Rehurek and P. Sojka, "Gensim–python framework for vector space modelling," *NLP Centre, Faculty of Informatics, Masaryk University, Brno, Czech Republic*, vol. 3, no. 2, 2011.

[37] M. Honnibal, I. Montani, S. Van Landeghem, and A. Boyd, "spaCy: Industrial-strength Natural Language Processing in Python," 2020.

[38] S. Bird, E. Klein, and E. Loper, *Natural language processing with Python: analyzing text with the natural language toolkit*. O'Reilly Media, Inc., 2009.

How Does the Language of 'Threat' Vary Across News Domains?

A Semi-Supervised Pipeline for Understanding Narrative Components in News Contexts

Igor Ryazanov[1] and Johanna Björklund[1]

***Abstract*— By identifying and characterising the narratives told in news media we can better understand political and societal processes. The problem is challenging from the perspective of natural language processing because it requires a combination of quantitative and qualitative methods. This paper reports on work in progress, which aims to build a human-in-the-loop pipeline for analysing how the variation of narrative themes across different domains, based on topic modelling and word embeddings. As an illustration, we study the language associated with the threat narrative in British news media.**

I. INTRODUCTION

Due to the drastic changes in news distribution over the past decades, considerable attention has been given to how ongoing events are framed in news reporting. In the realm of digital news media, concerns include increasing polarisation and a decrease in the relative share of political reporting [1]. News reporting has a direct effect on the political landscape because alternative news framing translates into competing public discourses and, by extension, electoral results [2]. Studying the framings and narratives in the media is, therefore, vital for understanding political processes. Extensive qualitative analysis of large-scale news corpora is, however, expensive and can hardly be feasible. This provides the motivation to apply natural language processing to both facilitate qualitative research at scale and enable quantitative approaches to narrative understanding. In this paper, we propose a pipeline for descriptive multi-domain analysis of narrative subcomponents in the news media.

Advancements in natural language processing (NLP) methods provide a variety of tools for political communication analysis. Some of these are found in applications within or adjacent to the area of narrative understanding, such as stance detection [3] and sentiment analysis [4]. Algorithms based on neural networks have been shown to discern the difference in published texts by different partisan actors [5] and predict the ideological alignment of social media posters [6]. The majority of these methods can be described as classification algorithms, relying on supervised learning on either narrow domain-specific annotated datasets or fine-tuning large language models (LLMs) which have been trained on huge general-domain corpora. They are primarily used for quantitative studies and practical applications where the target phenomena are well-defined and the domain shift is limited. The methods, however, are not always suitable to assist qualitative and descriptive research. The low level of interpretability of the machine learning methods in general, and of the LLMs in particular, is also a factor.

We are studying the usefulness of NLP for fine-grained narrative structure analysis, e.g. extracting narrative substructures and revealing context-specific language. The proposed pipeline, which is still a work in progress, serves to describe the contextual use of narrative themes within different overarching topics. As an example, we study the language used by news publishers to express the notion of threat and risk. We choose this type of semantic relation because its presence in a news article almost guarantees a degree of partiality which affects the readers' perception of the issue.

The pipeline consists of the following steps:

- Applying semi-supervised or unsupervised topic modelling to find latent topics in a text corpus
- Training contextual embeddings for each discovered topic
- Computing the closest terms in the embedding space to describe the notion of threat in each topic

In the system presented here, the embeddings are produced by Word2Vec, and topics are derived through Correlation Explanation (CorEx) where clusters are shaped around user-provided anchor words [7]. Depending on the available domain knowledge and discoveries from unsupervised clustering, the anchor words can guide the model to find crisper topics in a semi-supervised fashion. The output of the pipeline is a collection of descriptions of a selected concept (in our case, threat) for each of the generated topics.

II. BACKGROUND

In the context of NLP applications, interpretations and definitions of narrative structures and elements can vary greatly, depending on specific tasks and domains. Here we relate our problem to several of these approaches. While our goal to investigate the language abstract notions in different contexts (here exemplified by threat) does not match them exactly, it shares many similarities with e.g. stance detection and narrative discovery.

[1] I. Ryazanov and J. Björklund are with the Department of Computing Science, Faculty of Science and Technology, Umeå University, Umeå, Sweden igorr@cs.umu.se; johanna@cs.umu.se. This work is supported by Marianne and Marcus Wallenberg's Foundation, the Swedish Research Council, and WASP-HS.

A. Opinion mining and stance detection

As demonstrated in a number of studies, certain narrative-like notions can be captured by large language models trained on the document level. The documents or sentences are labelled by human annotators as containing such notions, and the definition of the notion is left to expert judgement. The assumption is that the representation of the document is rich enough that it captures narrative elements regardless of form. This is very prominent in, e.g. hate speech detection, where the key challenge comes from the fact that the hateful intent can take misleading forms and does not rely on any specific device to be conveyed [8], [9]. Similarly, it can be the case for the stance detection task, where the stance towards an issue cannot always be determined by positive or negative vocabulary and sentiment analysis is not enough to draw meaningful conclusions [3].

The downside is that the model trained to classify entire documents would be able to identify, e.g. a stance towards a specific political issue, but not necessarily what constitutes the narrative within the text. For example, an article can be shown to include the notion of 'threat', but explaining what constitutes 'threat' beyond the label becomes problematic. Since the algorithmic decision applies to an entire document, for narrative detection purposes, these models are more applicable to shorter messages and higher-level narratives ('pro-abortion' as opposed to 'threat' or 'success').

B. Narrative extraction

In the field of computational narratology, a common approach to narratives involves determining key entities and their relations. The inspiration for such methods comes from the structural interpretations of stories in formalist folklore studies [10]. Character or role detection can take various forms but often includes assigning a fixed set of archetypal roles or narrative frames ('villain', 'protagonist', 'victim', etc.) to specific entities in the story. In the news article domain, abstract role detection has been realised, for example, through the combination of entity extraction and sentiment analysis [11]. There has been, for example, some success in applying these methods in computational studies of conspiracy theories classifying entities within the publication as 'insiders' or 'outsiders' [12].

A more complex approach involves constructing the relations of the extracted entities, where the resulting narrative representation usually takes the form of a graph. This method has been applied, for instance, to conspiracy theory discovery. Relating this directly to our research, the authors used the presence of the notion of 'threat' encoded as subject-verb-copula triplets as the main criterion to detect specific theory elements [13]. We, however, are interested in how the narrative component of 'threat' is different in different news contexts and not in which contexts it defines.

C. Perspective extraction

Recently, Minnema et al. [14] put forth a framework based on Frame Semantics. They apply a FrameNet [15] parser LOME [16] to analyse perspectives in news media event description. Instead of building a graph, the focus is on analysing linguistic frames invoked by specific texts. While the purpose of the model is similar to our task, it is focused on the analysis of the specific events or topics (e.g. femicide reporting in Italy [17]) rather than comparing the contexts.

III. PRE-STUDY

A. Semi-supervised topic modelling for contextual 'threat' understanding

In our goal to keep the pipeline as robust and explainable as possible, we investigated the possibility of extracting contextual descriptions of 'threat' purely by applying semi-supervised topic modelling, which has the advantage of being interpretable in terms of probabilities, unlike Word2Vec word embeddings. Semi-supervised topic modelling has been used, for example, to investigate the presence of gendered latent topics in different contexts [18]. We explored the option of taking a similar approach, presupposing that there exists a specific cluster of news articles or their fragments centred around the target concept of threat. All of the pre-study has been performed on the same dataset as the rest of the paper and its thematical subsets: sports and politics. If the subsets would each contain a threat-related cluster of articles, we would have been able to compare their content and, therefore, the definitions of threat in these contexts.

B. Experiments with pSSLDA and CorEx

In the first series of experiments, we applied Latent Dirichlet Allocation (LDA) with z-labels (pSSLDA) [19]. At the initialisation step, it assigns additional weight to predefined seed words for specific topics. After initialisation, the algorithm proceeds in an unsupervised fashion. Through our experiments, we initialised clusters with various threat-related word combinations, as well as tested other similar notions, such as 'success'. The resulting topical distribution remained near-identical to the output of the unsupervised LDA model. Moreover, the topic order remained unstable even with the seeding, somewhat counterintuitively: one could have expected the military conflict-related news topic to be consistently initiated by seed words, such as 'danger'.

The second series of preliminary experiments using the more restrictive CorEx also displayed negative results: the topics initialised with the threat-related anchor (seed) words did not seem to be immediately humanly interpretable and had significant overlaps with other clusters. Our interpretation is that in a news dataset, the event-specific language dominates all other vocabulary particularities, making event-based topics very easily separable. So even if clusters of text corresponding to the notions, such as 'threat' exist, they remain statistically insignificant in comparison. While this may not be the case for other abstract topics, it is reasonable to expect a co-occurrence-based method to find clusters based on topic-specific terms rather than the presence of a higher-level semantic construct. Thus, we rejected using purely semi-supervised topic modelling for our task.

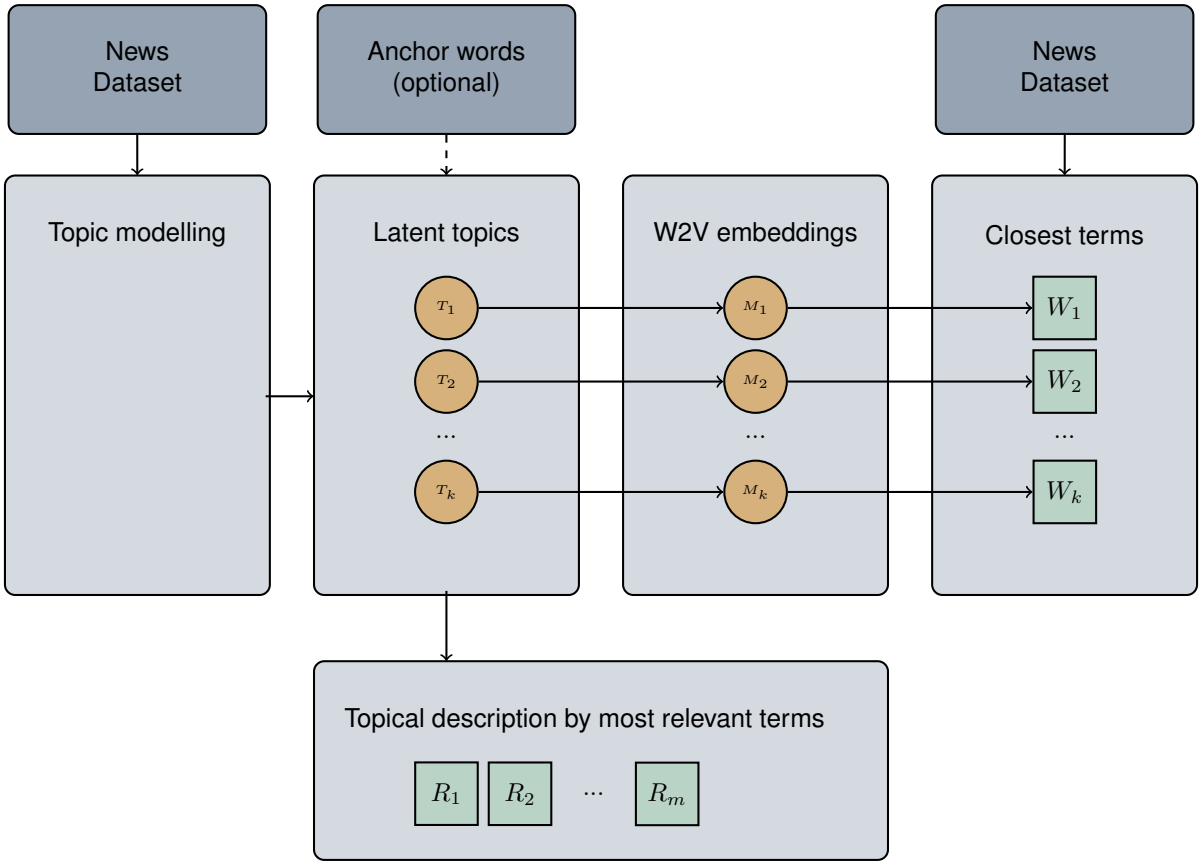

Fig. 1. In the proposed pipeline, the user provides a corpus for topic modelling, a parameter k declaring the expected number of topics, and, if desired, a number of anchor words to guide the topic formation. The result is k topics, represented by topical descriptions R_i consisting of the m terms most relevant to each. If the user is not satisfied, they may update the anchor words and repeat the clustering until the topics are as desired. In the next step, semantic embeddings M_i are computed for each topic T_i, and based on these, the closest set of closest terms W_i to the target concept are extracted.

IV. METHODS

A. Topic Modelling

To demonstrate the chosen approach, we apply it to study the language of threat in news media. The initial topic modelling is done with the help of CorEx, and in our analysis, we equate these topics with 'contexts'. Since the purpose of the pipeline is to perform exploratory analysis and assist in qualitative studies, it is vital to have some control over topic distribution based on domain knowledge. CorEx is based on mutual information between words and topics and offers more restrictive and flexible semi-supervised functionality compared to, e.g. Latent Dirichlet Allocation with z-labels [19]. It is also valuable for corpus exploration because it does not require transfer learning: Topic modelling based on neural networks can perform better on specific tasks [20], but also introduces additional bias from out-of-context corpora. Due to the nature of the models, this bias cannot be easily separated from the properties of the target datasets, which is a disadvantage in exploratory analysis.

B. Word Embeddings

After the topics have been established, we train Word2Vec embeddings on texts in each individual topic. Word embeddings preserve some degree of semantic relations from natural language [21] and have been used for investigating the definitions of concepts, as well as the narratives surrounding them. For example, Papasavva et al. [22] apply Word2Vec to find words associated with QAnon. It has also been used to compare semantic contexts: e.g. the language use of parliamentary motions of the opposing Swedish political parties [23]. While our goal is somewhat different, the approach is similar. We use a set of keywords as a representation for the target concept: in this case, threat.

Since our principal interest is investigating concrete media contexts, we avoid language models that require pre-training. Such models would inject implicit out-of-context bias, making comparisons harder. Additionally, the document cluster for each topic is relatively small, and it has been shown that Word2Vec can outperform transformer-based language models on smaller datasets when trained from scratch [24].

V. EXPERIMENTAL SETUP

A. Dataset

To evaluate the pipeline (outlined in Figure 1), we experiment with a collection of news articles from mainstream free-access British news media. The dataset was collected

between May and early August 2022 and contains 57,996 unique articles out of 100,000 in total. Top-5 most frequent news sources are *The Sun, The Independent, The Daily Express, The Daily Mirror and The Daily Star* together constituting approximately 23% of the articles in the dataset.

For each article, the following information was collected: the title (headline), the preamble, the body, the URL to the article, and the publication date and time. For the purposes of the experiments, the headline, the preamble and the body are concatenated into single text entries. The articles are tokenised and processed into the matrix of token counts with the Scikit-learn library.

B. Threat Definition

In this experiment, our goal is to define the notion of threat through the common language terms that encapsulate the relation of one entity presenting a threat to another. In that, our definition is similar to the definition of the frames in FrameNet: a frame is "a script-like conceptual structure that describes a particular type of situation, object, or event along with its participants and props" [15]. The semantically closest frame of FrameNet is 'Risky_situation' – "A particular Situation is likely (or unlikely) to result in a harmful event befalling an Asset.", so we choose to use the nouns in the FrameNet ontology that invoke the 'Risky_situation' (*threat*, *danger*, *risk*) as a triple of keywords.

VI. EXPERIMENTS

A. Unsupervised Topic Modelling

We assume minimal domain knowledge and task CorEx to identify news topics without anchor words to detect latent topics in the dataset. The number of topics is initially chosen to be below 10 to limit the scope of further analysis and after the preliminary experimentation set to 5. We evaluated the top-20 most relevant terms for each cluster, and in four cases out of five, the topics seem to have a clear focus (top-3 terms listed in parenthesis):

- T_0*: (league, season, premier)*
- T_1*: (government, cost, crisis)*
- T_2*: (police, court, officers)*
- T_3*: (love, instagram, star)*
- T_4*: (like, just, think)*

The final topic T_4 has the lowest topic correlation value. It is based on non-topic-specific words and seems to include articles not matching with the rest of the clusters.

B. Semi-supervised Topic Modelling

Based on the initial result above, we can expect that the four topics (loosely defined as 'Sports', 'Costs crisis' 'Police', and 'TV and celebrities') are present in the data, but it would be unreasonable to assume that all (or even most) articles would belong to one of them. Gallagher et al. [7], when investigating the performance of CorEx, set the number of clusters as high as 50 for a news dataset while initialising some of them with anchors to create crisp topics. The four topics described above are already shown to be present in the data. Even with minimal domain knowledge, we can also expect significant coverage of the Russia-Ukraine war in the summer of 2022, which is a potentially valuable context for threat interpretation. To isolate the five chosen topics, we restrict them with three anchor words each and set the total number of topics to 20.

The anchor words and the resulting topics are shown in Table I. Each of them is described by the top 10 most relevant terms. The list includes anchor words used for initialisation (in bold). Corex is a discriminative model and allows articles to belong to several topics.

C. Word Embeddings

For the articles in each topic, we train a Word2Vec model and extract the terms in the embedding space that are the closest to the three keywords: *threat*, *danger* and *risk*. We use the Gensim implementation of Word2Vec with the following parameters for all individual models: ignoring unique words (frequency 2 or more), word window size – 10, vector size – 100. The results are presented in Table II. For each context–keyword pair, we show seven words with the closest vector representation measured by cosine distance. Words unique to the contexts are highlighted in bold. As we can see, these neighbourhoods vary greatly between the contexts.

D. Analysis

At this step in the pipeline, the automatic analysis could be complemented with expert knowledge to add a qualitative element to the study. However, even by analysing the output of the models superficially and without specialised knowledge, we notice certain peculiarities. One such thing is that the language of the 'cost crisis' topic is as strong if not stronger than the language of the 'war' topic. Another observation is that the word 'fetus' is likely present together with 'court' because the Roe v. Wade ruling was overturned by the US Supreme Court within the time frame. With greater domain knowledge, one could choose better anchors and obtain crisper clusters. It is, for example, likely the term 'TV' caused the last topic to skew towards war instead of the cultural sphere as was intended.

VII. DISCUSSION

A. Limitations

One immediate limitation of the pipeline is the need for human guidance in the clustering process. While annotated data is not necessary, clustering does benefit greatly from domain knowledge, as the unsupervised version is unlikely to produce meaningful results. Another related disadvantage is the lack of 'ground truth' knowledge to test the results with the problem framed as it is. We, however, work on the assumption that a media researcher would not look for purely quantitative output in their application, using this framework as a facilitator for qualitative research instead.

Another potential weakness is the need to define any potential target concept with the keywords. While it is not in itself problematic with a domain expert's input, we so far lack the evidence to judge what kind of concept definition is preferable in a general case.

TABLE I

FIVE TOPICS INITIATED WITH ANCHOR WORDS FROM THE ORIGINAL DATASET.

Id	Size	Top-10 terms
T_0	4018	***war, ukraine, russia***, *russian, invasion, military, putin, ukrainian, forces, vladimir*
T_1	13,594	***league, player, sport***, *players, football, squad, goals, champions, clubs, winger*
T_2	8249	***cost, crisis, economy***, *living, struggling, cuts, poverty, spiralling, poorest, unemployment*
T_3	10,134	***police, court, officers***, *arrested, incident, investigation, crime, victim, judge, guilty*
T_4	8070	***tv, celebrity, singer***, *series, ekin, davide, su, island, luca, sanclimenti*

TABLE II

THREAT, RISK, AND DANGER IN FIVE CONTEXTS (IDENTIFIED LATENT TOPICS), REPRESENTED BY THE CLOSEST TERMS. TERMS UNIQUE TO EACH TOPIC IN THIS SELECTION ARE HIGHLIGHTED IN BOLD.

Context	Key word	Top-7 closest terms
'War'	*threat*	***geopolitical***, *escalation,* ***strategic***, ***warfare***, *threats,* ***counter***, *risks*
	danger	***context***, *grave,* ***disaster***, *dangers, height,* ***catastrophe***, ***breadth***
	risk	*risks,* ***situation***, *safety,* ***disaster***, *consequences, escalation, threat*
'Sports'	*threat*	***creativity***, ***backline***, ***pressing***, ***presence***, ***attack***, ***defence***, ***possession***
	danger	***trouble***, *threat,* ***territory***, ***control***, ***possession***, ***fall***, ***lines***
	risk	*risks, cause,* ***reduce***, ***potentially***, *size,* ***financial***, ***prevent***
'Cost crisis'	*threat*	*escalation, conflict, threats, grave, geopolitical, danger, nuclear*
	danger	*threat,* ***midst***, *fear,* ***prospect***, *consequences,* ***serious***, *concern*
	risk	*risks, damage,* ***serious***, *consequences,* ***expense***, *causes, cause*
'Police'	*threat*	***aggression***, *conflict,* ***wider***, ***issue***, *escalation,* ***terrorist***, ***outrage***
	danger	*risk,* ***unnecessary***, ***fetus***, *fear, cause,* ***distress***, *potentially, harm*
	risk	*risks,* ***potentially***, *danger,* ***effective***, *level, cause, levels, nature*
'TV & celebrities'	*threat*	*nuclear,* ***opposition***, ***economic***, ***missiles***, *conflict,* ***kremlin***, ***russia***
	danger	***circumstances***, ***saturated***, ***elements***, ***doom***, ***arteries***, ***cynicism***, ***turmoil***
	risk	*levels, height, safety,* ***consumers***, ***increases***, *damage,* ***assistance***

B. Future work

In our ongoing project, we have set ourselves several goals that would expand on existing experiments:

- We plan to study how the frequency and distribution of keywords or their combinations affect the results. As we propose to use the pipeline to study relatively high-level and not explicitly domain-specific concepts, we would like to at least outline how unspecific the keywords need to be. Based on this, we hope to provide a high-level recommendation on how to define the concepts. Similarly, we plan to formulate a recommendation on how to guide topic modelling with anchor words.
- Then, we aim to repeat the experimental scenario for other abstract narrative concepts, such as 'success' or 'failure', and compare the model's performances.
- The next step is to extend the experiments to an analogous dataset of the Swedish media. While it is reasonable to expect topic modelling and Word2Vec to work similarly well in another Germanic language, the news media culture is different, which is likely to cover the use of language.

Moving further, we can see this pipeline being used in comparative studies of news publishing language in different contexts, not only limiting them to event-based topics. The contexts can include, e.g. different types of publications (mainstream vs tabloids vs new media) or political alignments ('left wing' vs 'right wing'). Another potential use case is comparing the language of the same publication over a time period to investigate how language shifts within particular news contexts. Finally, an even more challenging task requiring more topical expertise would be drawing comparisons between the same concepts in the news media of different countries in their respective languages.

C. Conclusion

We have implemented a semi-supervised pipeline to analyse the expression of narrative themes in different media contexts. Previous studies have used word embeddings to describe terms and stances, and we extend this to a more abstract notion and produce a complete pipeline to perform comparative analysis. Our next steps include applying the pipeline to other languages and comparing the performance and results to English-language media. We also reach out to media researchers to identify other relevant applications. We believe that when there is sufficient domain knowledge to guide topic formation, this mixed-method approach can be an effective tool for narrative analysis.

ACKNOWLEDGEMENT

This work was supported by the Wallenberg Al, Autonomous Systems and Software Program – Humanities and Society (WASP-HS), as well as the Marianne and Marcus Wallenberg Foundation and the Swedish Research Council.

We are thankful to Hannah Devinney and Anton Eklund for their helpful comments on an early version of this manuscript that did much to improve the content. We are also grateful to Adlede AB for their support in the data collection.

REFERENCES

[1] P. Van Aelst, J. Strömbäck, T. Aalberg, F. Esser, C. De Vreese, J. Matthes, D. Hopmann, S. Salgado, N. Hubé, A. Stępińska, *et al.*, "Political communication in a high-choice media environment: a challenge for democracy?," *Annals of the International Communication Association*, vol. 41, no. 1, pp. 3–27, 2017.

[2] L. Alonso-Muñoz and A. Casero-Ripollés, "Populism against europe in social media: The eurosceptic discourse on twitter in spain, italy, france, and united kingdom during the campaign of the 2019 european parliament election," *Frontiers in communication*, vol. 5, p. 54, 2020.

[3] D. Küçük and F. Can, "Stance detection: A survey," *ACM Computing Surveys (CSUR)*, vol. 53, no. 1, pp. 1–37, 2020.

[4] M. Birjali, M. Kasri, and A. Beni-Hssane, "A comprehensive survey on sentiment analysis: Approaches, challenges and trends," *Knowledge-Based Systems*, vol. 226, p. 107134, 2021.

[5] T. Fagni and S. Cresci, "Fine-grained prediction of political leaning on social media with unsupervised deep learning," *Journal of Artificial Intelligence Research*, vol. 73, pp. 633–672, 2022.

[6] W. Chen, X. Zhang, T. Wang, B. Yang, and Y. Li, "Opinion-aware knowledge graph for political ideology detection.," in *IJCAI*, vol. 17, pp. 3647–3653, 2017.

[7] R. J. Gallagher, K. Reing, D. Kale, and G. Ver Steeg, "Anchored correlation explanation: Topic modeling with minimal domain knowledge," *Transactions of the Association for Computational Linguistics*, vol. 5, pp. 529–542, 2017.

[8] F. Poletto, V. Basile, M. Sanguinetti, C. Bosco, and V. Patti, "Resources and benchmark corpora for hate speech detection: a systematic review," *Language Resources and Evaluation*, vol. 55, pp. 477–523, 2021.

[9] A. Schmidt and M. Wiegand, "A survey on hate speech detection using natural language processing," in *Proceedings of the fifth international workshop on natural language processing for social media*, pp. 1–10, 2017.

[10] V. I. Propp, *Morphology of the Folktale*, vol. 9. University of Texas Press, 1968.

[11] D. Gomez-Zara, M. Boon, and L. Birnbaum, "Who is the hero, the villain, and the victim? detection of roles in news articles using natural language techniques," pp. 311–315, 2018.

[12] P. Holur, T. Wang, S. Shahsavari, T. Tangherlini, and V. Roychowdhury, "Which side are you on? Insider-Outsider classification in conspiracy-theoretic social media," pp. 4975–4987, 2022.

[13] S. Shahsavari, P. Holur, T. Wang, T. R. Tangherlini, and V. Roychowdhury, "Conspiracy in the time of corona: Automatic detection of emerging covid-19 conspiracy theories in social media and the news," *Journal of computational social science*, vol. 3, no. 2, pp. 279–317, 2020.

[14] G. Minnema, S. Gemelli, C. Zanchi, T. Caselli, and M. Nissim, "SocioFillmore: A Tool for Discovering Perspectives," pp. 240–250, 2022.

[15] C. F. Baker, C. J. Fillmore, and J. B. Lowe, "The berkeley framenet project," in *COLING 1998 Volume 1: The 17th International Conference on Computational Linguistics*, 1998.

[16] P. Xia, G. Qin, S. Vashishtha, Y. Chen, T. Chen, C. May, C. Harman, K. Rawlins, A. S. White, and B. Van Durme, "LOME: Large Ontology Multilingual Extraction," pp. 149–159, 2021.

[17] G. Minnema, S. Gemelli, C. Zanchi, V. Patti, T. Caselli, and M. Nissim, "Frame semantics for social nlp in italian: Analyzing responsibility framing in femicide news reports.," in *Italian Conference on Computational Linguistics*, 2021.

[18] H. Devinney, J. Björklund, and H. Björklund, "Semi-supervised topic modeling for gender bias discovery in english and swedish," in *GeBNLP2020, COLING'2020–The 28th International Conference on Computational Linguistics, December 8-13, 2020, Online*, pp. 79–92, Association for Computational Linguistics, 2020.

[19] D. Andrzejewski and X. Zhu, "Latent dirichlet allocation with topic-in-set knowledge," in *Proceedings of the NAACL HLT 2009 Workshop on Semi-Supervised Learning for Natural Language Processing*, pp. 43–48, 2009.

[20] M. R. Grootendorst, "Bertopic: Neural topic modeling with a class-based tf-idf procedure," *ArXiv*, vol. abs/2203.05794, 2022.

[21] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," *Advances in neural information processing systems*, vol. 26, 2013.

[22] A. Papasavva, J. Blackburn, G. Stringhini, S. Zannettou, and E. D. Cristofaro, ""is it a qoincidence?": An exploratory study of QAnon on Voat," in *Proceedings of the Web Conference 2021*, pp. 460–471, 2021.

[23] A. Fredén, M. Johansson, P. Kisic Merino, and D. Saynova, *A Comparison of Language Processing Models in Political Analysis: Evidence from Sweden*. Oct. 2021.

[24] A. Edwards, J. Camacho-Collados, H. De Ribaupierre, and A. Preece, "Go simple and pre-train on domain-specific corpora: On the role of training data for text classification," in *Proceedings of the 28th international conference on computational linguistics*, pp. 5522–5529, 2020.

