POLARIS
TECHNICAL REVIEW
INNOVATION JOURNAL

LEONARDO LABS

A glance to new perspective in advanced research

48-2023

LEONARDO

PROPRIETARY NOTICE

contents

editorial

‘Thinking today about tomorrow’s factory’. This is what Leonardo is pursuing with its Leonardo Labs programme, to which the POLARIS Innovation Journal dedicates this second issue that completes the presentation of some of the research carried out within the Labs and the medium-term results that the company expects to achieve with this initiative.

Some articles stem from the exploration of enabling solutions that are closer to the market and more oriented towards solving contingent problems. Conversely, other papers aim at results that are expected over a longer timeframe, over medium to long-term time horizons. These latter are typical to advanced research areas, which must be not exposed to the pressures of productivity and competition, nor to the compression of the long timeframes that are inherent to research.

Different approaches with different time scales coexist in our Labs, in which the Research arises from Universities and Innovation Centres combines with in-house research, to produce those positive effects - disruptive by definition - that are expected to open new perspectives, some of which are still difficult to imagine, by now. This is the dual approach of the Leonardo Labs which, although operating in different fields, are based on transversal approaches and are strongly interconnected in networks.

Thanks to the interdisciplinary approach and the Labs’ broad freedom of action, this is the combination of factors that allows us to imagine and investigate those solutions that do not exist yet, which we must be able to master in our future. It is also a matter of exploring emerging technologies, to exploit advantages and spin-offs that this kind of research, which must adapt to different needs, has on both the design and the technological upgrading of products.

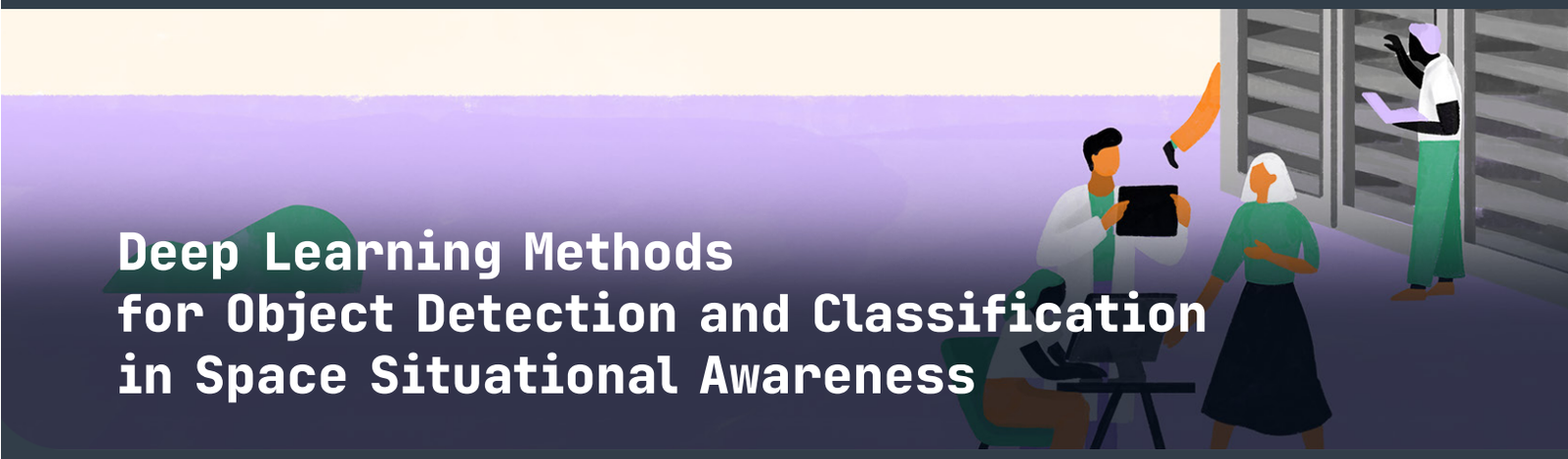
Some of these investigative perspectives are anticipated by the articles included in this issue, which focus on Artificial Intelligence and Deep Learning applied to Space and hyperspectral images for detection of the objects on Earth (remote sensing, geology, environmental monitoring, and target detection) or in Space itself (space debris). Artificial Intelligence and High Performance Computing (HPC) as enabling technologies in research, are the focus of further articles dedicated to New Materials and Supercomputing for real time storage and processing of large masses of data.

Then, the frontier of frontiers: the application of quantum principles to computing, sensor technology and precision computing. It is very important to understand and master these new paradigms. In our vision, this is a commodity to be drawn upon. Of course, our goal is not to deliver a quantum computer, but to study applications of quantum technologies. These go ranging from use of the entanglement to realise sensors capable of hither to unimaginable performances, to quantum clocking that will make satellite platforms ever more precise, to quantum communication based on the distribution of cyber-attack resistant keys.

In setting up its Labs, Leonardo ‘has bet’ on the talents of Innovation. The research activities carried out by those young people who joined the company over the past two years, will find their concrete expression in a future that at this time we are not interested in knowing how near or far it could be. In fact, the idea of the Leonardo Labs arises, even more than from operational requirements, from that feeling shared throughout Leonardo, which animates and pervades the project of the Labs in itself and includes also the investment on the talents of Innovation, which is where our actual ‘bet’ lies.

To those young researchers, who represent our future, we have decided to give trust and means which enable their mind to go sailing the unpredictable environment of the knowledge economy.

Editor in Chief
Vincenzo Sabbatino



Deep Learning Methods for Object Detection and Classification in Space Situational Awareness

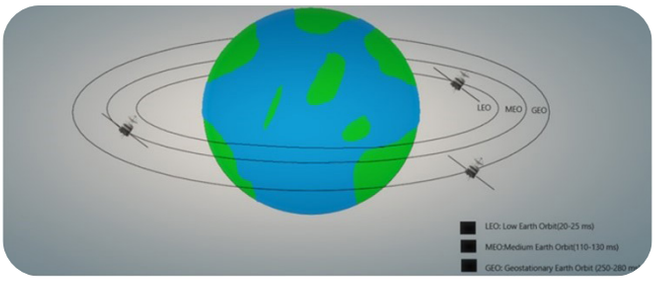
Federica Massimi[1], Pasquale Ferrara[2], Francesco Benedetto[1]

[1]Università Roma Tre - Signal Processing for Telecommunications and Economics Lab,
[2]Leonardo Labs – Future Electronics and Sensing

The number of space objects, mega-constellations of satellites and debris in Low Earth Orbit, is continuously increasing. The growing risk of space collisions poses a significant threat to sustainability and safety of the space operations, which must be carefully and efficiently addressed to avoid critical damages to the satellite networks. This paper aims at providing a survey about the state-of-art, challenges, and perspectives regarding the use of deep learning methods for Space Situational Awareness. Furthermore, we present a case-study to demonstrate the benefits of using deep learning for space object detection by radar observations.

INTRODUCTION

Among all of Earth's orbital regimes, the Low Earth Orbit (LEO), generally defined as the region in space between ~160 to 2,000 km in altitude (Figure 1), is by far the most congested because of its closeness to Earth [1]. Over the last decade, companies are placing satellites into such orbit at an unprecedented frequency, to build mega-constellations of communications satellites.



1 – Different types of satellite orbits around the Earth and the round-trip latencies

These systems, with even tens of thousands satellites in LEO, are nowadays becoming a reality [2]. Many studies have been carried out especially for low orbit satellites, due to their interesting characteristics, such as low latency and large capacity [3].

As a consequence, the number of active and not anymore operative satellites in LEO has increased by over 50%, to about 5000. Space X alone is on track to add 11,000 more units as it builds its Starlink mega-constellation. Others have similar plans, including One Web, Amazon, Telesat, and GW, a Chinese state-owned company. Thousands of satellites and rocket bodies provide considerable mass in LEO, which can break into debris upon collisions, explosions, or degradation. Fragmentation of satellites increases the cross-section of orbiting material, and with it, the collision probability per time.

There is much debris or "space junk" in the near-Earth space environment that is large enough to threaten human spaceflights and robotic missions, but too small to be tracked.

This debris population includes both natural meteoroid and man-made orbital debris, called "orbitals".

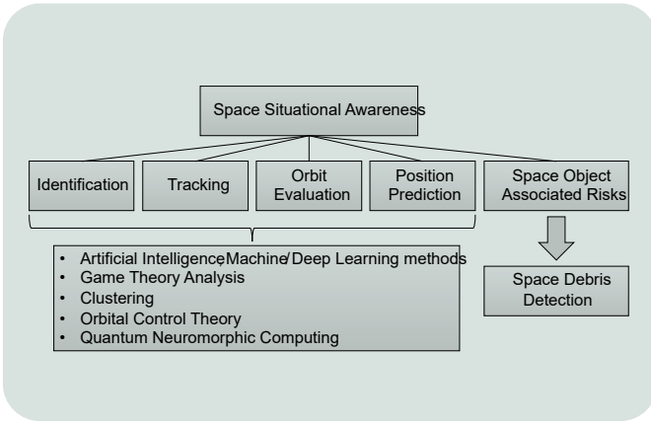
Orbital debris generally are any man-made object in orbit around the Earth that no longer serves a useful function, such as non-functioning spacecraft, abandoned launch vehicle stages, mission-related debris, and fragmentation debris. They are constantly tracked every day, and more than 27,000 (Space Surveillance Networks estimate) [4] pieces have been detected so far. Since both debris and spacecraft travel at extremely high speeds (about 25,267 km/h),

(about 25,267 km/h), the impact of even a tiny piece of debris with a spacecraft could create dramatic problems. There are half a million pieces of debris the size of a marble or larger (up to 1 centimetre), and about 100 million pieces of debris about 1 millimetre and larger. There is also even smaller micrometre-sized debris and about 23,000 pieces of debris larger than a softball orbiting the Earth. Even small flecks of paint can damage a spacecraft when traveling at those speeds. Simulations of the long-term evolution of debris suggest that LEO is already in the protracted initial stages of the Kessler Syndrome [5], and this could be managed through active debris removal [6].

While large commercial satellite constellations undeniably offer tremendous potential for the satellite industry, they inevitably increase the probability of mutual collisions among orbiting objects. This poses a significant threat to sustainability and safety of the space operations, which must be carefully and efficiently addressed.

The operation of monitoring the space environment and resident space objects, by identifying and characterizing space objects and their operational environment is known as Space Situational Awareness (SSA) [7].

More in details, the SSA focuses on problems related with: (i) space objects tracking, (ii) their identification, (iii) determining their orbits, (iv) gaining knowledge about the scenario in which they are operating, and (v) forecasting their upcoming positions and related risks to their functioning (Figure 2).



2 – SSA objectives and enabling technologies

The SSA is hence of fundamental importance to all the space traffic management operations, as one of the main aspects of the SSA is to calculate debris from fragmentation events, meteor storms, or other natural events that may be very dangerous for all the space systems, and to act consequently. One critical task is to classify space objects according to their properties.

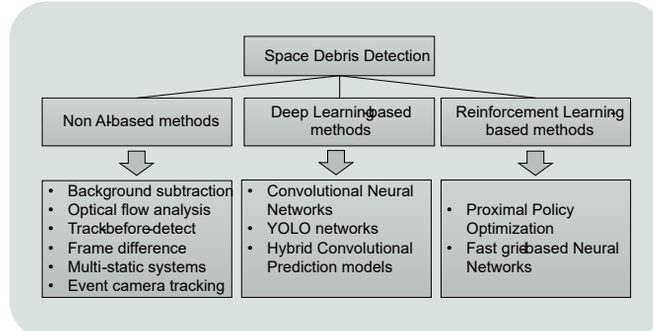
Unfortunately, the information available about space objects is often limited. Typically, the visual magnitude and the radar cross section of space objects are obtained via optical and radar sensors, respectively. Artificial Intelligence (AI) and machine learning (ML) systems appear very promising to detect and classify this kind of objects. Methods reported in literature include Artificial Neural Networks [8], reinforcement learning [9], and deep neural networks [10] to assess behavioural analysis [11] and autonomy [12]. These ML methods, as well as deep learning (DL) methods [13], support evidence-based knowledge for Space Domain Awareness [14] programs such as the DARPA Hallmark. The SSA also includes the understanding of mission policies, technical aims, and orbital mechanics [15]. In addition, the SSA benefits from game-theory studies in supporting pursuit-evasion analysis [16], and from gathering data to track satellites, debris, and natural phenomena [17]. Then, in order to get effective SSA results, tracking efforts must coordinate with detection policies [18], waveform selection [19], and attack mitigation [20].

New trends are focusing on agile, intelligent, and efficient computer vision architectures, operating on quantum neuromorphic computing, as part of the SSA network [21]. The quantum neuromorphic vision, combined with polarimetric dynamic vision sensors principles, represents the SSA of tomorrow, as it works at very high speeds, with poor requirements about bandwidth, power, and memory.

In the next section, we focus on the space object detection, since orbital debris pose the highest end-of-mission risk for most spacecraft roboticists operating in LEO. For an exhaustive review of the state of the art, please refer to [22].

LITERATURE REVIEW

As it is shown in Figure 3, the literature on space object detection can be categorized in: (i) Non AI-based, (ii) Deep learning-based, and (iii) Reinforcement learning-based methods.



3 – Space debris detection: enabling technologies

Non AI-based methods

This class of algorithms can be divided into:

- Background subtraction methods [23] to compare the moving parts (foreground) of a video with a background image;
- Optical flow analysis methods [24] to calculate motion vectors for the points within the images and it estimates where the points might be in the next image;
- Track-before-detect methods [25]: usually, target detection is performed before tracking. In this paradigm, sensor data on a provisional target are integrated over time, and the target is detected without the use of any threshold;
- Frame difference methods [26] to compare the difference between two consecutive video frames at pixel level. An interesting example of a tracking system based on camera for events is presented in [27].

Most of these techniques work well in almost any ideal condition (e.g., little blur and low noise). Recently, in [28], authors have evaluated the pros and cons of using either a multi-static radar or telescopes for SSA. The advantage of using optical telescopes is that they enable observing objects at much longer distances compared to radar observations. On the other hand, radar systems can operate 24/7 regardless of weather and light (i.e., sunlight) conditions. Radar can fine-tune the transmitted signal to perform proper processing and estimate the physical and dynamic properties of the target, providing the highly accurate measurements that are essential for orbit determination. However, one of the major disadvantages of using radar sensors for SSA purposes is their high cost, as such systems require very high transmitting power

Deep learning-based methods

Artificial Intelligence can be very useful for processing and optimizing the large amount of data collected by scientific missions such as space probes, Earth observation spacecraft and rovers. AI also enables a variety of monitoring tasks, thanks to ubiquitous satellite imagery in space. The Deep Learning (DL), as a subset of AI, can enable precise and automated control and facilitate onboard activities, such as docking or navigation [29]. Compared to traditional techniques, the DL techniques require little or even none pre-processing steps for training. Thus, features that may be not identifiable to a human can be automatically and efficiently extracted. Furthermore, DL architectures are suitable for modelling complex behaviours of multimodal dataset, and most of the time they outperform traditional techniques in terms of accuracy.

Convolutional networks have widely proven their worth in the areas of image recognition and classification. Images go through convolutional layers aiming to extract different aspects or features of the image, and to assign them weights and biases to classify them. In [30], the authors propose a method for detecting salience of the space debris, which is based on a fully convolutional network for the space surveillance platform. Meantime, the network directly learns the internal relationship between two frames, thus avoiding expensive optical flow calculations. However, such method is not suitable for detecting small targets, due to the limited features it extracts from small space debris. In [31], a novel U-Net deep neural network is exploited for image segmentation and real-time extraction of tracklets from optical acquisitions. A new type of convolutional network is presented in [32]. PSnet is a perspective sensitive network to detect objects from different perspectives (i.e., angles of view). The features are mapped to the pre-set multi-perspective spaces to obtain the specific semantic feature of the object, decoupled from the angle of view.

The work [33] proposes an end-to-end space-craft image segmentation network that combines the DeepLabv3+ semantic segmentation with a multiscale neural network based on sparse convolutions with attention mechanism. Experiments on a spacecraft segmentation dataset show that the encoder-plus-attention-plus-encoder structure can obtain clear and complete masks of spacecraft targets, with high accuracy.

The authors in [34] explore deep neural networks applied to light curve data and compare the performance of classical classification algorithm, CNN and Recurrent Neural Networks (RNN).

Finally, YOLO models are used in many real-time object detection applications. In [35], YOLO is used to rank regions of interest, found by using background subtraction. Optimized architectures, such as the YOLOv3-Tiny [36], have been also used for human targets detection from on-board unmanned aerial vehicles (UAVs).

Reinforcement learning-based methods

Reinforcement Learning is the area of AI wherein agents are trained to take decisions or actions by interacting with the environment. In [37], authors simulated a controllable ground telescope observing satellites in LEO by using the Double Deep Q-Learning. Once the number of satellites observed over a period is maximized, they used an extended Kalman filter to significantly reduce uncertainties of the kinematic measurements.

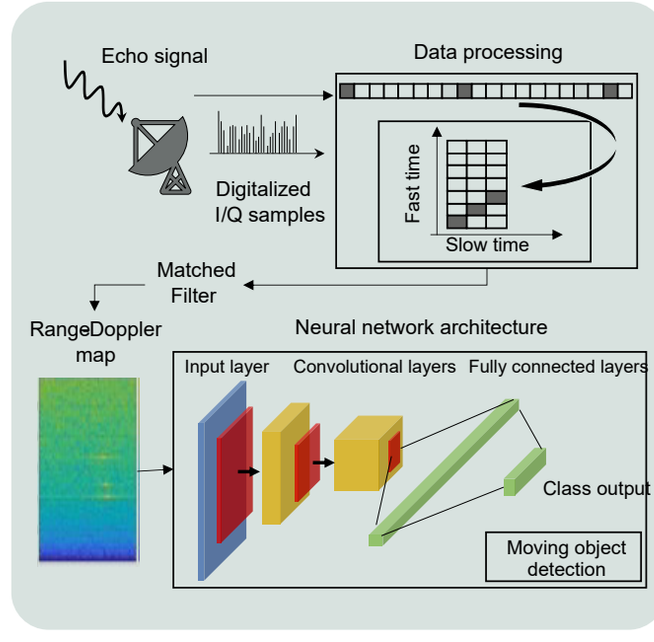
Devices such as Floating Space Manipulators (FFSM) are increasingly used in various space activities and the active object tracking is the basis of many space missions. These systems present two major challenges: the modelling and control of FFSM and the planning of the tracking motion. To address them, the paper [38] presents a strategy for active object tracking of the FFSM by employing the deep reinforcement learning (DRL), the proximal policy optimization, and a fuzzy neural network.
DRL does offer a fresh perspective on how to handle challenging space missions, and it works well in conjunction with more conventional approaches. Data acquisition costs, training efficiency, sensitivity to parameters, and environment are just some of the open challenges. In [39], four DRL agents are trained/tested on a simulated SSA environment that can support arbitrary sensor position, various resident space objects, observation windows and sensor properties (action change and stabilization time, dwell time, measurement models and process noises).
Both the fully connected and the CNN agents are explored, as well as their resilience to changes in orbital regimes, observation window lengths, observer positions, and sensor change rates.
The agents have shown solidity to most of these variations and continue to outperform short-sighted policies.
Sensor tasking and sensor management problems have been addressed in [40]: the asynchronous advantage actor-critic (A3C) model combines the benefits of both the policy-based and the action-based approaches, to get more accurate policy gradient estimations and to enhance convergence. This approach can also pick up information asynchronously from a collection of agents that engage with the environment. Although some preliminary findings were presented, results still need to be explored further.
Finally, in [41] authors proposed a fast detection method of space debris with grid-based learning. The image is divided into 14×14 grids, then a fast grid-based neural network is used to pinpoint the location of the spatial debris in the grids. Such method demonstrates excellent accuracy (98.8%) and high detection speed by processing an image every 2.3 ms.

CASE STUDY

Some preliminary experiments have been carried out by applying deep learning methods for space object detection in a simulated environment.
In our tests we focus on a well-defined and paradigmatic use-case i.e., the small moving object detection from radar signals. To do that, we simulate the processing chain of a monostatic pulse-Doppler radar that detects the radial velocity of moving targets at specific ranges.
The radar output is used to feed a neural network architecture that provides the number of detected targets, as shown in Figure 4.

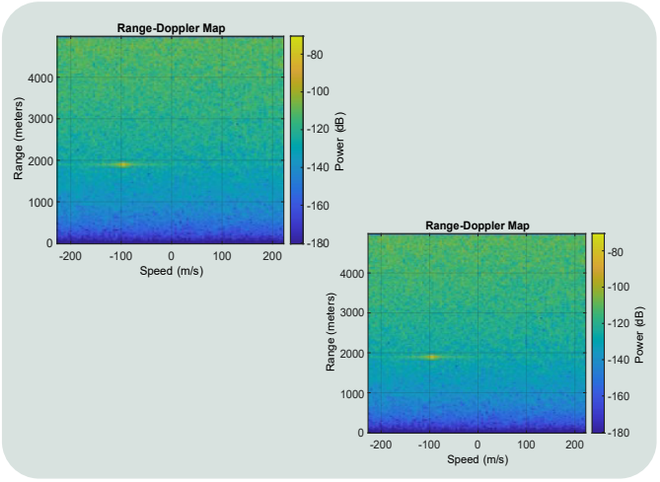
Radar Signal Processing

At the receiver, the electromagnetic echoes reflected by a target are split into in-phase (I) and quadrature (Q) components by using coherent de-modulation. As a propagation form, we assume the two-rays ground reflection model, while targets are characterized in terms of their distance (i.e. range) from the radar, velocity, and Radar Cross Section.



4 – Flowchart of the experiments

Pulse-Doppler signal processing separates the reflected I/Q signals into channels, by means of a set of filters for each ambiguous range. Afterwards, such samples are reshaped into a time domain matrix, whose columns correspond to range samples (fast time), while its rows correspond to pulse intervals (slow time).
Once convolved with the matched filter by means of a Fourier transform, the output matrix provides a power spectral density estimate of the returned signal, in function of the range and Doppler frequency. This matrix is also known as range-Doppler map. In a classical pulse-Doppler processing chain, the estimation of target position and velocity is performed by thresholding the range-Doppler map and by finding the range and the Doppler bins in which the energy exceeds the given threshold. In our experiments, these maps are the input of the neural network architectures under test.



5 – Maps with one (left) and three (right) moving objects

Neural Network frameworks

The architectures that have been tested are described in the following.
Squeeze Net is a very light architecture that nevertheless achieves out-standing performance in computer vision tasks. Squeeze Net is a small neural network, with a few parameters, which can easily adapt to portable devices thus having lower computational and memory demand to reduced inference time. It is made up of 18 layers. Unlike in most networks, the last learnable layer is the final convolutional layer.
VGG-16 is a 16-layer deep neural network with about 138 million parameters. For this reason, it takes a long time to process input data, as well as it occupies a significant amount of memory. VGG-16 is composed of a stack of 13 convolutional layers followed by three fully connected layers: the first two layers have 4096 nodes each, while the last layer has been adapted to our use-case (i.e., 4 output nodes). In our test, we performed a new training by freezing the first ten layers and by learning new weights for the other six layers.
Alex Net consists of 5 convolutional layers, followed by 3 fully connected layers. It also allows for parallel-GPU training by splitting its neurons on different devices. Dropout layers within the first two fully connected layers are used to avoid over-fitting, at the price of increased time for the model convergence.
GoogLe Net architecture is 22 levels deep with 9 starter modules stacked linearly. The first convolutional level uses large multi-dimensional filter kernels to reduce the number of input channels and to avoid overfitting, without losing spatial information.

Results and discussions

Network	Precision	Recall	F1_score
Squeeze Net	93.7%	94.0%	93.7 %
VGG-16	87.5 %	88.5 %	87.2 %
Google Net	85.2 %	85.5 %	84.2 %
Alex Net	92.5 %	93.2 %	92.8 %

Table 1 – Performance of the tested deep neural networks

Experiments have been conducted on a dataset of range-Doppler maps, generated by randomly varying the objects position (from 1 to 3000 m) and speed (from -225 to 225 m/s). We have assigned to each map one of the 4 possible labels, each of them representing the number of targets detected (0 if there are no objects, 1 if there is only one target and so on). Once generated, the range-Doppler maps (see Figure 5) have been passed to the neural networks, to classify whether the image does belong to the first, second, third or fourth case mentioned. The entire dataset consists of 800 images: 80% of the maps were used for the training phase, while 20% for validation. We have used the Stochastic Gradient Descent with Momentum optimizer, and the

learning rate is 0.0003 with 50 epochs.
The classification performance is measured in terms of precision, i.e., TP/(TP+FP), and recall, i.e., TP/(TP+FN), where TP, TN, FP, and FN are the number of true positives, true negatives, false positives, and false negatives, respectively. Precision quantifies the number of positive class predictions that belong to the positive class, while recall is the number of positive class predictions obtained from all the positive examples in the dataset. Their harmonic mean, called F1_score, provides a single score. The results are reported in Table 1.
The Squeeze Net has demonstrated to be the best performing network in terms of overall F1_score. The Alex Net network follows in terms of performances. Using fewer convolutional layers has the advantage of featuring lower hardware needs and shorter training times than the VGG-16 and the GoogLe Net.
One of the key design choices of the VGG-16 network is that it uses smaller convolutional filters than those of the previous networks.
This allows accurate recognition of the images (with an F1_score of 87.2%), at the cost of dramatically increasing the depth of the network. The Google Net network has turned out to be the least performing (84.2% of F1_score).

CONCLUSIONS

Space infrastructures are subject to collisions with debris, abandoned space objects and other active satellites, every day. This is a big issue, as modern society relies heavily on space infrastructure for operations such as communications, guidance and navigation, weather forecasting and spatial images.
Therefore, it is crucial to be aware of the space situation and to develop new algorithms for the defence of space infrastructures. This paper focused on such dramatic issues, by reviewing the recent papers published in the field of machine and deep learning for object detection in Space Situational Awareness systems. We also conducted experiments for space object detection by simulating the use of a monostatic pulse radar.
Results of such simulations demonstrate the efficiency of such methods for moving space objects detection, thus potentially improving collisions avoidance with space debris.
The endless overpopulation of the LEO will call for new required performance, a rising number of sensors, and a growing need for SSA data acquisition.
New research should invest in optimization of the system's understanding capacity, as well as of the multisource data fusion.
Then, several important shortfalls must be counteracted, such as the need for large training datasets, and for very long training times. Looking ahead, the need to work with short monitoring intervals that would enable reacting quickly to debris decomposition for satellite manoeuvres, will be a hot research topic area, thus improving the capacity of non-stop orbital prediction, tracking and monitoring in hazardous situations.
Hence, research will be conducted in order to strengthen the confidence of the system, by improving early warning collision systems as well as studying more accurate manoeuvring avoidance policies.

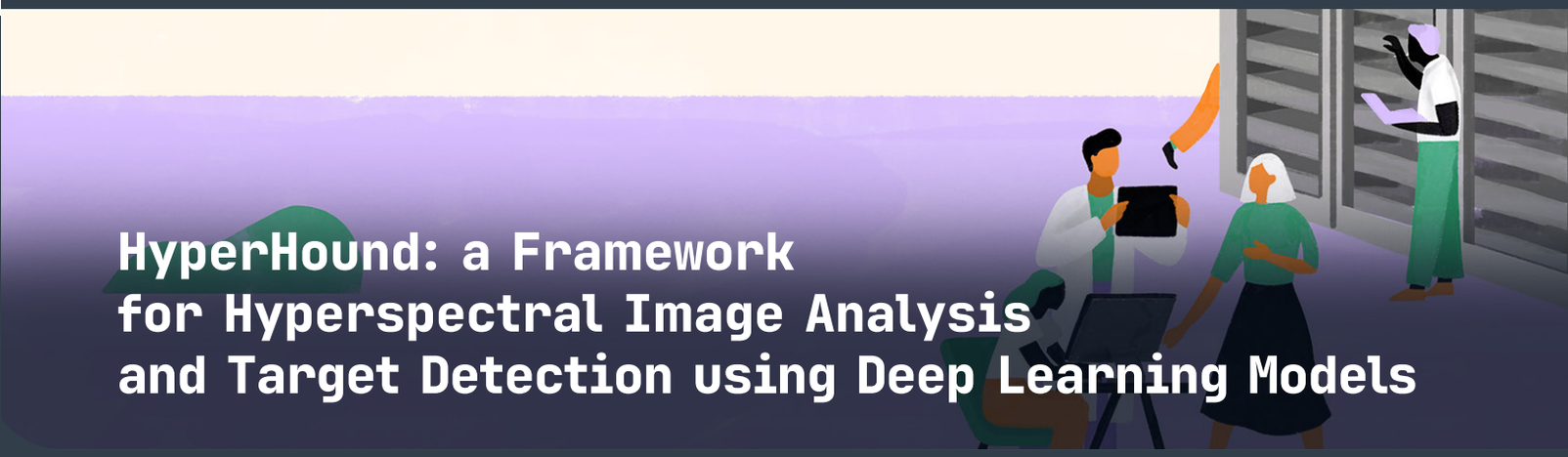
Pasquale Ferrara: pasquale.ferrara02.ext@leonardo.com

REFERENCES

[1] M. Dominguez et al., "Space traffic management: assessment of the feasibility, expected effectiveness, and funding implications of a transfer of space traffic management functions", National Academy of Public Administration, Washington, DC, USA, Aug. 2020.
[2] M.A. Sturza, and G.S. Carretero, "Mega-Constellations – A Holistic Approach to Debris Aspects", in Proc. 8th European Conference on Space Debris, Darmstadt, Germany, 20–23 April 2021.
[3] Tasneem Darwish et al., "Location Management in Internet Protocol-Based Future LEO Satellite Networks: A Review", IEEE Open Journal of the Communications Society, vol.3, pp.1035-1062, 2022.

[4] ESA. Space Environment Statistics: Space debris by the numbers. 2021 Last updated: 2021-01-08. https://sdup.esoc.esa.int/discosweb/statistics/.
[5] D. Kessler, and B. Cour-Palais, "Collision frequency of artificial satellites: The creation of a debris belt", J. Geophys. Res., vol. 83, 2637, 1978.
[6] J.-C. Liou, and N.L. Johnson, "Risks in space from orbiting debris", Science, vol. 311, 5759, 2006.
[7] B. Jia et al., "Space Object Classification Using Deep Neural Networks", in Proc. of IEEE Aerospace Conf, 2018.
[8] H. Peng and X. Bai, "Artificial Neural Network-Based Machine Learning Approach to Improve Orbit Prediction Accuracy", Journal of Spacecraft and Rockets, vol. 55, no. 5, pp. 1248-1260, 2018.
[9] R. Linares and R. Furfaro, "Dynamic Sensor Tasking for Space Situational Awareness via Reinforcement Learning", Adv. Maui Optical and Space Surveillance Techn. Cont. (AMOS), 2016.
[10] B. Jia et al., "Space Object Classification Using Deep Neural Networks", IEEE Aerospace Conf., 2018.
[11] R. Furfaro et al., "Resident Space Object Characterization and Behavior Understanding via Machine Learning and Ontology-based Bayesian Networks", Advanced Maui Optical and Space Surveillance Tech. Conf., 2016.
[12] J. Valasek, Advances in Computational Intelligence and Autonomy for Aerospace Systems, AIAA, 2019.
[13] U. Majumder et al., Deep Learning for Radar and Communications Automatic Target Recognition, Artech House, 2020.
[14] A.D. Jaunzemis et al., "Evidence-based sensor tasking for space domain awareness", AMOS Tech, 2016.
[15] D. Shen et al., "Network survivability-oriented Markov games (NSOMG) in wideband satellite communications", IEEE/AIAA Digital Avionics Systems Conference, 2014.
[16] D. Shen et al., "Pursuit-Evasion Games with Information Uncertainties for Elusive Orbital Maneuver and Space Object Tracking", Proc. SPIE 9469, 2015.
[17] Z. Hall and P. Singla, "Reachability Analysis Based Tracking: Applications to Non-cooperative Space Object Tracking", 3rd Int. Conf. on Dynamic Data Driven Applications Systems, 2020.
[18] Y. Ding et al., "Blind Transmission and Detection Designs with Unique Identification and Full Diversity for Noncoherent Two-Way Relay Networks", IEEE Trans. Veh. Technol., vol. 63, pp. 3137-3146, Sep. 2014.
[19] Z. Shu et al., "Game theoretic power allocation and waveform selection for satellite communications", Proc. SPIE 9469, 2015.
[20] C.T. Do et al., "Game Theory for Cyber Security and Privacy", ACM Computing Surveys, vol. 50, no. 2, pp. 30, 2017.
[21] C. Barnes et al., "Space Situational Awareness (SSA) and Quantum Neuromorphic Computing," 2022 IEEE International Conference on Imaging Systems and Techniques (IST), 2022, pp. 1-6.
[22] F. Massimi et al., "Deep Learning Methods for Space Situational Awareness in Mega-Constellations Satellite-Based Internet of Things Networks", Sensors 2023, 23, 124.
[23] R. Kalsotra, S. Arora, "Background subtraction for moving object detection: explorations of recent developments and challenges", Vis Comput, 2021.
[24] D.H. Diamond et al. "Accuracy evaluation of sub-pixel structural vibration measurements through optical flow analysis of a video sequence", Measurement, Volume 95, 2017, Pages 166-172, ISSN 0263-2241.
[25] W. Yi et al., "Multi-Frame Track-Before-Detect Algorithm for Maneuvering Target Tracking," in IEEE Transactions on Vehicular Technology, vol. 69, no. 4, pp. 4104-4118, April 2020.
[26] A.M. Husein et al., "Motion detect application with frame difference method on a surveillance camera", Journal of Physics: Conference Series, 2019.
[27] K. Xiao et al., "A Preliminary Research on Space Situational Awareness Based on Event Cameras," 2022 13th International Conference on Mechanical and Aerospace Engineering (ICMAE), 2022, pp. 390-395.
[28] D. Cataldo, et al., "Multibistatic Radar for Space Surveillance and Tracking," in IEEE Aerospace and Electronic Systems Magazine, vol. 35, no. 8, pp. 14-30, 1, 2020.
[29] V. Kothari et al., "The Final Frontier: Deep Learning in Space", arXiv, 2020.
[30] J. Tao et al., "Deep Convolutional Neural Network Based Small Space Debris Saliency Detection," 2019 25th International Conference on Automation and Computing (ICAC), 2019, pp. 1-6.
[31] A. De Vittori et al." Real-time space object tracklet extraction from telescope survey images with machine learning". Astrodyn 6, 205–218, 2022.
[32] X. Zhang et al., "PSNet: Perspective-sensitive convolutional network for object detection", Neurocomputing, Volume 468, 2022, pp. 384-395, ISSN 0925-2312.

[33] Y. Liu et al., “Multi-Scale Deep Neural Network Based on Dilated Convolution for Spacecraft Image Segmentation”, Sensors, 2022, 22, 4222.
[34] B. Jia et al., “Space object classification using deep neural networks,” IEEE Aerospace Conference, 2018, pp. 1-8.
[35] C. Kim et al., “A hybrid framework combining background subtraction and deep neural networks for rapid person detection”, J Big Data, 5, 22, 2018.
[36] G. Liu et al., “The Development of a UAV Target Tracking System Based on YOLOv3-Tiny Object Detection Algorithm,” 2021 IEEE International Conference on Robotics and Biomimetics (ROBIO), 2021, pp. 1636-1641.
[37] B. Oakes et al. “Double Deep Q Networks for Sensor Management in Space Situational Awareness,” 25th International Conference on Information Fusion, 2022, pp. 1-6.
[38] W. Lei et al., “Active object tracking of free-floating space manipulators based on deep reinforcement learning”, Advances in Space Research, 2022, ISSN 0273-11177.
[39] P.M. Siew and R. Linares, “Optimal Tasking of Ground-Based Sensors for Space Situational Awareness Using Deep Reinforcement Learning”, Sensors, 2022, 22, 7847.
[40] R. Linares and R. Furfaro, “An Autonomous Sensor Tasking Approach for Large Scale Space”, University of Arizona
[41] Y. Xiang et al., “Space debris detection with fast grid-based learning,” IEEE 3rd International Conference of Safe Production and Informatization, 2020, pp. 205-209.



HyperHound: a Framework for Hyperspectral Image Analysis and Target Detection using Deep Learning Models

Rosario Di Carlo, Roberto Morelli, Matteo Guidi, Alessandro Nicolosi

Leonardo Labs - Applied Artificial Intelligence

Hyperspectral images have shown great potential for the target detection task. These images collect the reflectance physical value over a large electromagnetic spectrum providing a fingerprint that characterizes uniquely distinct materials. In this work, a framework is developed to recognize different materials using several approaches ranging from classical methods to Deep learning ones. Different learning paradigms are investigated considering both supervised and metric learning methods. The main difference between these approaches concerns the labelling process. Indeed, while the former method requires labelling the data, the latter approach is based on pseudo-labels generation described in this contribution.

INTRODUCTION

Hyperspectral imaging (HSI) [1] is an advanced technology that allows the collection of a wide range of spectral data acquired by remote sensors. It has been shown to be useful for various applications, including object detection, classification, and material recognition. In particular, hyperspectral images provide unique material fingerprints that can be used to identify different materials.

In recent years, there has been an increasing interest in developing machine learning models that can accurately recognize materials from hyperspectral images.

Deep learning has emerged as a promising approach to solving complex problems in various fields. Among the different deep learning models, convolutional neural networks (CNNs) [2] have become dominant for processing visual-related tasks. CNNs are a class of biologically inspired multilayer neural networks that can be trained end-to-end from raw image pixel values to classifier outputs. The concept of CNNs was first introduced in a paper by LeCun et al.[3] and has since been improved upon by subsequent research [4] and refined and simplified by other studies [5][6].

In this paper we propose a framework that leverages both the classical and deep learning approaches for material recognition in hyperspectral images. We investigate different learning paradigms, including supervised and self-supervised methods, and evaluate their performance on a benchmark dataset.

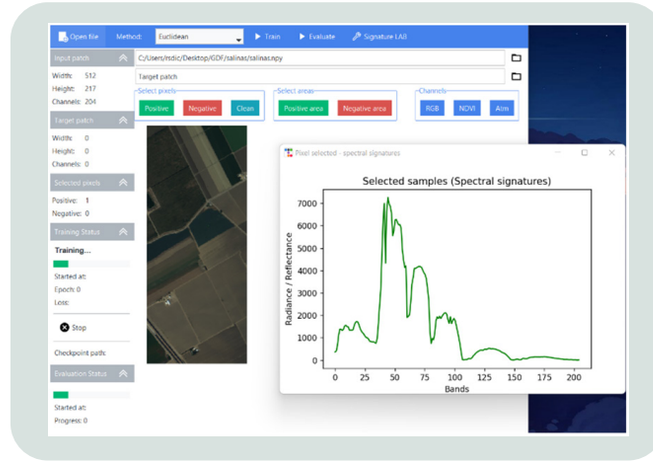
Our approach shows promising results and can have practical applications in fields such as remote sensing, geology, environmental monitoring, and target detection.

HYPERHOUND FRAMEWORK

HyperHound is a framework developed specifically for analysing hyperspectral images. It has been designed to allow an easy implementation and testing of various models for target detection.

This framework comes with a broad range of capabilities, with its main features described below. All these features are provided with a simple user interface (UI) shown in Figure 1.

Compatibility with the PIX format: support loading files in PCI (Geomatics Database File) format, splitting them into smaller patches, and visualizing them for the analysis process.
Datasets: Integration of both publicly available and privately collected datasets to enable model evaluation and comprehensive data analysis.
Implementation of classic target detection models: target detection is a critical task in computer vision, which involves identifying specific objects of interest within an image, HyperHound implements several algorithms that provide a solid foundation for measuring the performance of newer and more advanced models. Implementing these classic models allows us to compare the results of different models and evaluate their relative strengths and weaknesses. Some of the classical models implemented are Euclidean distance, CEM, MF, and ACE.
Data labelling: the interface of HyperHound provides two options for labelling data, individual pixel labelling and bounding box selection. The labelled data can be used to train a classification model.
Functionalities of Inference and Training: HyperHound implements functionalities to perform both inference with pre-trained deep learning models and train models on the fly from the interface. The inference process is optimized by splitting the input image into smaller slices and processing them in parallel on a GPU.



1 - UI of Hyperhound framework loading Salinas data and analysing a spectral signature of the selected pixel

Database of spectral signatures: consisting of laboratory-sampled materials collected from online sources. This resource enables comparisons between the reflectance of individual pixels and available materials, enabling the computation of similarity scores.
Atmospheric correction: Integration with py6s [7], a Python implementation of the 6S model [8], to compute atmospheric correction of the spectral image according to the atmospheric conditions during the data acquisition.

METHODS

Standard Methods

Classical hyperspectral image target detection algorithms, such as Spectral Angle Mapper (SAM) [9] and Spectral Information Divergence (SID) [10] are two straightforward detection algorithms that measure the “distance” between the spectrum of the test pixel and the prior spectral signature of the target.
Also Constrained Energy Minimization (CEM) [11][12] matched filter (MF) [13], and adaptive coherence/cosine estimator (ACE) [14][15] are typically developed using constrained least square regression methods or hypothesis testing methods that assume a Gaussian distribution. However, real-world hyperspectral data obtained through remote sensing often exhibit strong nonlinearity and non Gaussianity, which can result in a decline in the performance of these classical detection algorithms.

Standard Methods

Self-supervised learning is a type of machine-learning technique in which a model is trained to learn patterns and relationships within a dataset without the need for explicit labelling or supervision. For the scope of this work, this method is used to learn a space topology to cluster similar hyperspectral signatures. In this sense, starting from a reference signature, this algorithm can detect similar targets from the images analysed. To overcome the labelling burden, an unsupervised method is used to generate pseudo-labels.

The strategy used in this work is described in the evaluation and results section and leverages a clustering pre-text task. Once pseudo-labels are generated, contrastive learning is used to train the model to cluster properly signatures belonging to distinct classes. It is worth remembering that these are the classes defined in a self-supervised manner, that is, using an unsupervised pre-text task. A fully connected neural network was chosen to learn the distance metric for class discrimination.

Fully-connected neural network (FCNN)

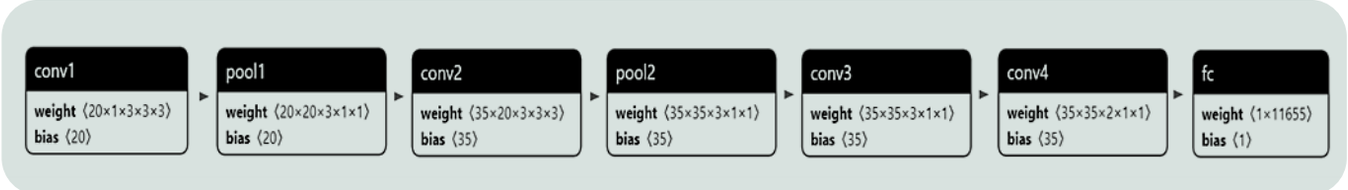
A fully-connected neural network consists of a series of fully connected layers that connect every neuron in one layer to every neuron in the other layer.
Each neuron represents a computational unit that processes its input and passes its results to each neuron of the next layer. Layer by layer a hierarchical representation of the input is learned to improve the classification task that consists of producing a probability for each pixel to belong to the target object. For the hyperspectral images, the input of the FCNN is represented by all the channels of a single image pixel that are processed consequently by all the fully-connected layers. Indeed, the first layer of the network has an input dimension equal to the hyperspectral channels while the other layers have a number of neurons that gradually decreases. The last layer has a number of neurons equal to the dimension of the code used to encode the pixel given in the input.
Indeed, the network is trained to encode the input into a sequence of numbers in a latent space. In this way, pixels belonging to the same class are clustered together to reduce their distance into the latent space. To promote this behaviour, the training proceeds by means of a metric learning approach as explained at the beginning of this paragraph.

Supervised

The supervised learning method involves training a model using labelled training data, which consists of a set of inputs and their corresponding outputs or class labels. The model's parameters are updated iteratively during the training phase to accurately predict the desired outputs. In the testing phase, the model is evaluated against new input or test data to assess its ability to predict the correct labels. With sufficient training, the model can predict the labels of new input data. However, this approach requires a large amount of labelled training data to fine-tune the model parameters. Therefore, it is most appropriate for situations where much-labelled data is available. The HyperHound framework facilitates this labelling process and the following training. The model adopted to test the framework is a convolutional neural network with 3D convolutional filters.

3D Convolutional neural network (3D-CNN)

Identifying ground objects in hyperspectral imaging requires both spectral and spatial information. To effectively classify these objects, a 3D convolutional neural network (CNN) was implemented. The network processes each pixel of the images by considering the relation between adjacent channels, in addition to spatial patterns across neighbouring pixels. The input of the 3D-CNN is a patch of 7x7xN pixels, where N is the number of channels in the hyperspectral image. The architecture consists of a series of 3D convolutional layers, with decreasing filter numbers leading to the last fully connected layer.
This final layer takes the flattened concatenation of a set of feature maps from the last convolutional layers as input and outputs the probability of the centre pixel of the input patch belonging to a target object. A scheme of this neural network architecture is reported in Figure 2.



2 – CNN3D Architecture

Hyperparameters optimization

The HyperHound framework included this architecture after extensive hyperparameter optimization to find the best model in terms of validation loss, which is related to the detection performance of the model. To scale our search for optimizing hyperparameters, we employed Ray Tune, a Python library designed for executing experiments and tuning hyperparameters at any scale. This was done using the Leonardo HPC system, specifically the davinci-1 infrastructure, which comprises a total of 80 nodes, each equipped with four Nvidia A100 GPUs.

EVALUATION AND RESULTS

In the following are described first the datasets used for the self-supervised approach and his evaluation. Following the same scheme is then introduced also the dataset used for the supervised method together with the relative labelling process and the performance assessment.

Salinas

Is a hyperspectral dataset collected by the 224-band AVIRIS sensor over Salinas Valley, California, and is characterized by high spatial resolution (3.7-meter pixels). The area covered comprises 512 lines by 217 samples. 20 water absorption bands were discarded: [108-112], [154-167], 224 for a total number of bands equal to 224. This image was available only as at-sensor radiance data. It includes vegetables, bare soils, and vineyard fields. Salinas Ground-Truth (GT) contains 16 classes.

Data Labelling

The self-supervised approach for labelling involves generating samples that are labelled without full supervision. One method for accomplishing this is through the use of endmembers, which are defined as pure spectral signatures that can be linearly combined to represent the hyperspectral image pixels. Endmembers can be thought as the basis vectors of a geometrical subspace. During image acquisition, due to the relatively low spatial resolution of hyperspectral sensors, some pixels may collect a mix of signatures from different materials. This means that each pixel can be seen as a superimposition of each endmember. By identifying the endmembers in the hyperspectral image, it is possible to obtain a set of pure spectral signatures that can be used to label the image data. Once the endmembers have been identified, they can be used in a variety of ways to label the image data. For example, one approach is to use spectral unmixing to estimate the abundance fractions of each endmember in each pixel.

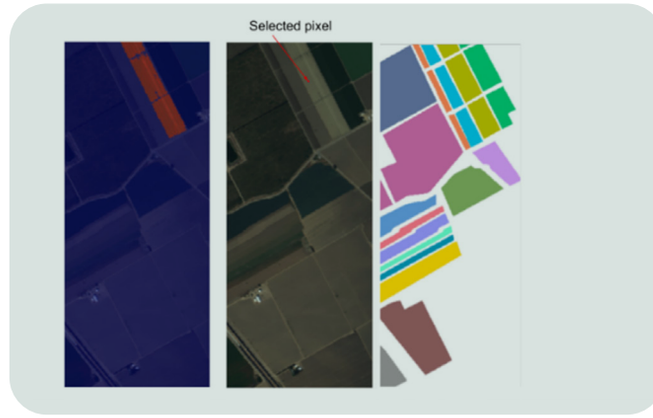
Nevertheless, some methods exist to unmix the pixel to find the basic constituents of each material, but their application doesn't guarantee the optimality of the solution. Indeed, the method used should define both the exact number of pure signatures inside a picture and the relative abundances of each end member that represent the coefficient of the linear mixing. Both these parameters are unknown and so the solutions are ill-defined. However, a guess about the number of endmembers is made to perform the unmixing. Once the endmembers are defined, the dataset generation can be provided by sampling the coefficients that define their linear mixing following the equation:

$$x = \sum_{i}^{n} e_i * c_i \quad (1)$$

where x is the new generated sample, c_i are the randomly generated coefficients and e_i represents the n endmembers extracted from the source image. This sampling is repeated to generate all the dataset samples. The key step in this process is the labelling step, where the endmember corresponding to the highest coefficient is used as the label y_i. In other words:

$$y_i = Argmax(c_i) \quad (2)$$

So, in the end, a dataset with a custom number of samples is generated with several classes equal to the number of endmembers.

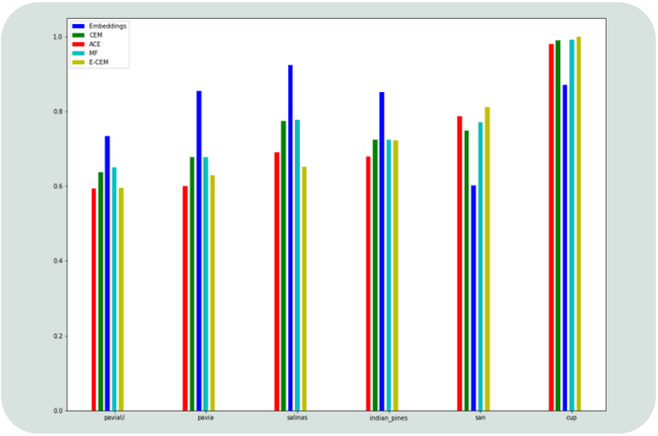


3 - Evaluation on Salinas dataset using only one selected target pixel belonging to the class "Stubble". Heatmap of the classification on the left, Selected target pixel on the centre, GT on the right

Performance

All the models were evaluated on different hyperspectral datasets, each containing one or multiple classes to detect. For each dataset, one of the classes was designated as the target class, while the others were considered background classes. This process was repeated for all the classes.

To identify the target class, a representative pixel was selected and the distance between that pixel and all other pixels in the dataset was computed. If the similarity score between the target pixel and another pixel was above a certain threshold, that pixel was classified as a target pixel; otherwise, it was classified as a background pixel.
This process was repeated for each class, with the threshold value modified to calculate the relative area under the curve (AUC) performances. The final AUC value was calculated by taking the mean value over all the classes. Using this approach, the ability of different models to detect target objects in hyperspectral datasets with multiple classes can be accurately measured. Figure 3 shows the inference results of the trained model, which was able to successfully detect the "Stubble" class. Figure 4 shows the results on 6 datasets: Pavia University, Pavia Centre, Salinas, Indian Pines, San Diego, Cuprite. It can be observed that the self-supervised model outperforms other models on most of the tested datasets.
We would also highlight that the datasets used in this study are composed of a single image with pixel-based labelling, which results in a scarcity of data variability.



4 – Comparison of the mean AUC on different datasets, Embeddings in blue refer to the self-supervised model

This can lead to a high correlation between the training and validation sets, which can negatively impact the evaluation of model robustness. This limitation can be observed in the paragraph below evaluating the same models on data acquired with real-world conditions variability.

Proprietary dataset

The dataset used for the supervised task is a proprietary dataset. It consists of 4 images collected with a hyperspectral sensor during an aerial acquisition. The images were pre-processed by performing the L1 pre-processing chain, which consists of the following operations:

- Spectral and Radiometric Calibration;
- Geo-Referencing;
- Geo-Rectification.

Data	Train	Validation	Test
Tiles containing the target (613x613)	11		7
Patches (7x7)	1050 (70%)	450 (30%)	Full tiles

Table 1 – Dataset training, validation, test splitting

These operations are missing the L2 pre-processing chain, which involves atmospheric correction and conversion of values to reflectance. In addition, the images have artifacts probably due to the vibrations the sensor was subjected to during flight.

Given these limitations, a single pixel may contain a mixture of multiple hyperspectral signatures. This constraint makes it difficult to discriminate small targets in the scene, leading to a decrease in the accuracy of the results.

Data Labelling

Each of the 4 images was cropped into 200 smaller tiles, each measuring 613x613 in size, for a total of 800 tiles. Through ground surveys, it was determined that 18 of these tiles contained the targets to identify. Of these 18 tiles, 11 were included in the training-validation sets, while the remaining 7 tiles were used to test the models. The labelling process is provided using the HyperHound interface. Through this interface, it is possible to display an image and collect a set of pixels to represent both target and background samples.

This collection can be performed by using both bounding boxes or dot annotations, for a finer pixel selection. This procedure was repeated on all 18 tiles used for the training, validation, and test. A patch of dimension 7x7 was cropped around each pixel collected to provide the input in the form of images to the 3D CNN used for the training. The total number of patches collected for training and validation was nearly 1500 with a proportion of 1:14 between target and background samples. The partition of data into training, validation and testing is summarized in Table 1.

Performance

As our objective is to identify target areas, we used a detection metric to evaluate the performance of the model. The F_1 score was chosen as the evaluation metric as it handles class imbalances better than other metrics such as accuracy. To assess the performance of the model, we developed an algorithm to associate the model's predictions with the ground-truth labels. The output of the model is a heatmap representing the probability of a pixel belonging to a target area. Therefore, the first step was to apply a threshold to obtain a binary mask, where each cluster of fully-connected pixels represents a predicted object. Subsequently, if a predicted object partially or fully overlaps with an object in the ground-truth mask, it is considered a true positive (TP). On the other hand, if there is no overlap between a predicted object and a ground-truth object, it is considered a false positive (FP). Finally, if a ground-truth label is not associated with any predicted object, the false negative (FN) count is increased by one unit.

$$F_1 = \frac{2 \cdot TP}{2 \cdot TP + FP + FN} \quad (3)$$

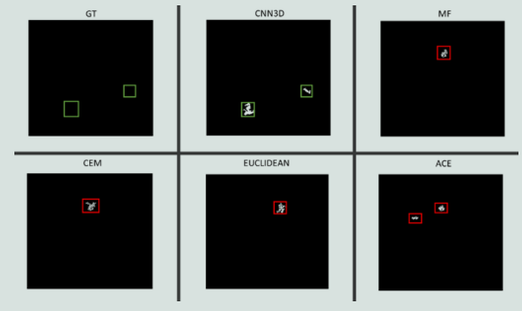
The results are provided in the Table 2.

The proposed model was evaluated on seven images that were not included in the training or validation set and achieved an F_1-score of 0.6 in identifying the targets. Notably, none of the other models tested were able to detect any of these targets.

An example of detection comparison on a test image is reported in in Figure 5. The first patch (top left corner) represents the ground truth, that is, a completely black image with green boxes corresponding to the targets to detect. The original image relative to this test is not shown to preserve sensitive information. The remaining patches represent the predictions of all the competing methods. Notably, only the CNN3D was able to detect correctly all the targets.

Table 2 – Detection results: number of true positives, false positives, false negatives

Tile n°	TP	FP	FN
1	2	0	0
2	2	2	0
3	1	0	0
4	0	3	2
5	1	1	0
6	1	2	0
7	1	1	0



5 – Comparison between our model (CNN3D) and the other standard methods

CONCLUSIONS

This article presents the HyperHound framework, which has been developed for hyperspectral image analysis. The framework provides an effective solution for analysing hyperspectral data by applying deep learning techniques. Two types of deep learning models were analysed using the framework: self-supervised and supervised. The self-supervised approach is particularly useful in addressing the challenges of a lack of labelled data and the difficulty of pixel-level ground truth annotation. The model learns to predict features from the input data itself, without any explicit supervision. This approach is particularly effective when the ground truth data is not available, and it has shown good results in many literature datasets. However, the self-supervised models are less robust and their detection metrics are generally lower compared with supervised models. The supervised model, on the other hand, utilizes ground truth data to train the model.

This type of model yields good results, even on real-world data, where classical and unsupervised models often fail. The supervised model type is particularly useful in cases where the ground truth is available and can be used to train the model. These models are robust and can provide accurate results even under different real-world conditions. In this study, it is highlighted that many hyperspectral datasets used as benchmarks lack sufficient data, and the training and validation data are often highly correlated, resulting in models that are not robust to different real-world conditions.
However, the supervised models have shown significant improvement and are particularly useful for man-in-the-loop applications. They provide an excellent tool for guiding and facilitating the task of an expert analyst in identifying targets, which is a challenging task in hyperspectral data analysis. Therefore, the HyperHound framework and supervised models provide a promising direction for hyperspectral data analysis, and they hold great potential for addressing the challenges of real-world applications.

FUTURE WORK

In our future work, we plan to incorporate functionalities that take into account atmospheric conditions into our models and expand the range of models that can be used with HyperHound.

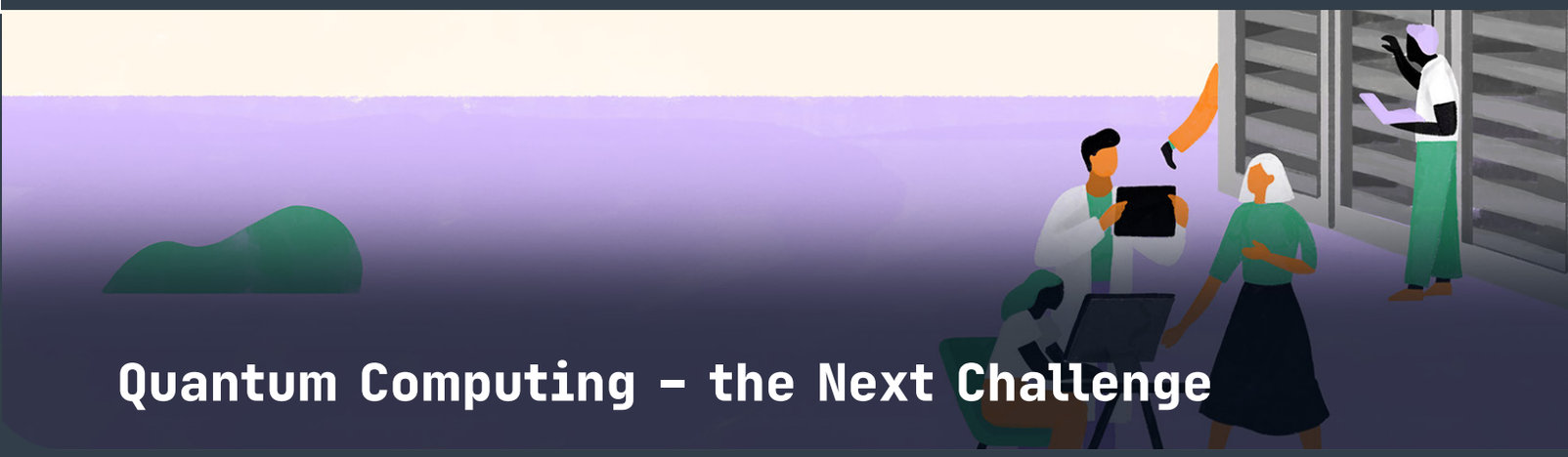
We will investigate the most promising methods for this task, such as the quantum-based [16] classifiers and compare them with our baseline models. Additionally, we will explore robust classification methods to improve the accuracy and reliability of our models. Specifically, we want to address the unbalanced problem that affects frequently such kinds of a dataset, where we have a lot of background samples and few target instances. We aim to test different kinds of loss, such as focal loss, and also many oversampling strategies that can help to improve detection accuracy.

Rosario Di Carlo: rosario.dicarlo.ext@leonardo.com

REFERENCES

[1] D. Landgrebe, “Hyperspectral image data analysis,” IEEE Signal Processing Magazine, vol. 19, no. 1, pp. 17–28, 2002.
[2] G. E. Hinton and R. R. Salakhutdinov, “Reducing the dimensionality of data with neural networks,” Science, vol. 313, no. 5786, pp. 504–507, 2006.
[3] K. Fukushima, “Neocognitron: a hierarchical neural network capable of visual pattern recognition,” Neural Networks, vol. 1, no. 2, pp. 119–130, 1988.
[4] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, “Gradient-based learning applied to document recognition,” Proceedings of the IEEE, vol. 86, no. 11, pp. 2278–2323, 1998.
[5] D. C. Cires¸an, U. Meier, J. Masci, L. M. Gambardella, and J. Schmidhuber, “Flexible, high performance convolutional neural networks for image classification,” in Proceedings of the 22nd International Joint Conference on Artificial Intelligence (IJCAI ’11), vol. 22, pp. 1237–1242, July 2011.
[6] P. Y. Simard, D. Steinkraus, and J. C. Platt, “Best practices for convolutional neural networks applied to visual document analysis,” in Proceedings of the 7th International Conference on Document Analysis and Recognition, vol. 2, pp. 958–963, IEEE Computer Society, Edinburgh, UK, August 2003.
[7] Wilson, R. T., 2013, Py6S: A Python interface to the 6S radiative transfer model, Computers and Geosciences, 51, p166-171
[8] Vermote, E. F., Tanré, D., Deuze, J. L., Herman, M., & Morcette, J. J. (1997). Second simulation of the satellite signal in the solar spectrum, 6S: An overview. IEEE transactions on geoscience and remote sensing, 35(3), 675-686.
[9] Kruse, F. A., A. B. Lefkoff, J. B. Boardman, K. B. Heidebrecht, A. T. Shapiro, P. J. Barloon, and A. F. H. Goetz. “The Spectral Image Processing System (SIPS) - Interactive Visualization and Analysis of Imaging spectrometer Data.” Remote Sensing of Environment 44 (1993): 145-163.

[10] Du, H., C.-I. Chang, H. Ren, F. M. D'Amico, and J. O. Jensen, J. "New Hyperspectral Discrimination Measure for Spectral Characterization." Optical Engineering 43, No. 8 (2004): 1777-1786.
[11] L. Gao et al., "Adjusted spectral matched filter for target detection in hyperspectral imagery," Remote Sens. 7(6), 6611–6634 (2015).
[12] Y. Cohen et al., "Subpixel hyperspectral target detection using local spectral and spatial information," J. Appl. Remote Sens. 6(1), 063508 (2012).
[13] D. Manolakis, "Detection algorithms for hyperspectral imaging applications," IEEE Signal Process Mag. 19(1), 29–43 (2002)
[14] E. J. Kelly et al., "Adaptive detection and parameter estimation for multidimensional signal models," NASA STI/Recon Technical Report 89, 28678 (1989).
[15] X. Jin et al., "A comparative study of target detection algorithms for hyperspectral imagery," Proc. SPIE 7334, 73341W (2009)
[16] D. Dragoni et al., "Quantum Computing – the Next Challenge", POLARIS Innovation Journal - Issue 48, 2023



Quantum Computing – the Next Challenge

Daniele Dragoni[1], Matteo Vandelli[1], Emanuele Triuzzi[2], Riccardo Mengoni[2], Daniele Ottaviani[2], Carlo Cavazzoni[1]

[1]Leonardo Labs - HPC/Cloud/Big Data Technologies, [2]CINECA Quantum Computing Lab

Quantum computing is a potentially disruptive computational paradigm that has the power to impact and rebuild the traditional technological system and to drive a new cycle of industrial revolution and transformation. While additional research and development are necessary, the existing technologies have shown promising results. In this paper we provide insights into the key trends in quantum computing and discuss how Leonardo is responding to the challenges posed by such technology. We provide a preview of the Leonardo Labs overall quantum computing research strategy, which is devoted to a pragmatic investigation of quantum algorithms at the bridge of the quantum computing and high-performance computing domains. Our initiatives aim to build strong internal competencies to enable effective and rapid adoption of this technology as it matures, while concurrently developing quantum-inspired, HPC-ready computational tools to enhance the company's digital competitiveness in the near term. Two lighthouse projects are hence presented, focusing on quantum machine learning and combinatorial optimization, which showcase how current quantum approaches can be applied to solve small-scale industrial problems.

INTRODUCTION

Quantum computing is a potentially disruptive computational paradigm that leverages the quantum mechanical principles of *superposition, interference*, and *entanglement* to process information stored in fundamental units known as qubits. Quantum computers (QCs) can carry out selected computational tasks that the classical digital hardware, even the most powerful supercomputers, cannot handle in a practical amount of time [1]. To realize this advantage, however, QCs necessitate dedicated quantum algorithms. Currently, various quantum algorithms exist that display a provable theoretical advantage with respect to the best-in-class classical counterparts for selected problem classes [2], offering computational speedups that grow even exponentially along with the problem size.

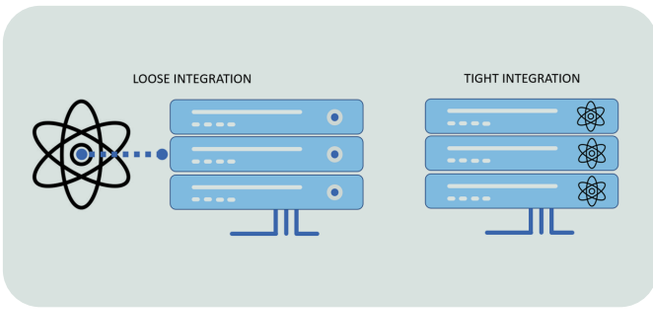
The most notable exponential speedup is provided by Shor's algorithm for the factorization of semi-prime numbers [3], which would enable breaking of the RSA encryption keys currently used for secure transactions [4] if a few thousands of ideal qubits were available.

Hype aside, however, the reality is that the current largest universal quantum computer features just 433 qubits [5], and, even if various quantum-hardware manufacturers have ambitious plans to construct devices with 1000+ qubits within a few years [6], it will take some time to enable the execution of speedup-proven algorithms for problem sizes that are of practical utility. The primary obstacle in scaling up the size of QCs is that quantity, quality, and connectivity of physical qubits must all grow at the same rate, which is challenging since errors induced by environment

disturbances and crosstalk increase quickly with the qubit count. Researchers are currently working to improve the hardware and to develop error-correction schemes based on qubit redundancy to alleviate this problem [7], but further work is required in this direction.
Currently, there is common consensus on the idea that future developments will unleash the full potential of fault-tolerant large-scale QCs [8]. Although this is hot expected to happen before five to ten years, this belief has triggered a competitive global race to achieve a first application that can solve real-world problems, and organizations have already started to explore how to harness the power of available noisy intermediate-scale quantum (NISQ) hardware[3].
Such NISQ devices typically feature 50-500 qubits, cannot execute fault-tolerant quantum algorithms with a provable speedup on practical scales, but have provided glimpses of quantum capabilities that go beyond those achievable with the most powerful high-performance computing (HPC) clusters. This has been clearly shown through the claims of quantum supremacy by organizations such as Google [9], Chinese research institutes[10], and Xanadu [11]. The competition is fierce and it is of paramount importance to enter the race as the earliest adopters, to be ready to exploit timely and effectively the technology as it will start creating value.

HPC & QUANTUM COMPUTING

Demonstrations of quantum supremacy have been achieved on mainframe NISQ devices only in purely academic problems. Nevertheless, NISQ computers need to be supplemented with classical digital hardware to deal with real-world problems. This classical support is currently implemented by means of a *loose-integration* paradigm, in which application workloads are divided between classical resources, typically local machines with consumer capacities, and quantum resources usually accessed remotely. In this precompetitive era of experimentation, the sharing of sensitive or classified data is not of a particular concern, and this approach is considered the most effective. Indeed, pragmatically, it enables testing a variety of technological solutions that are still scarce and geographically sparse (to date only a few countries host QCs), also providing enough agility to mitigate the exposure to those platforms that will inevitably suffer from rapid obsolescence.



1 – Loose integration vs Tight integration QC-HPC paradigms

A *tight-integration* paradigm is however also under scrutiny for future implementations. This paradigm involves designing QCs to operate *on-premises* and to be physically co-located with traditional computing resources, serving as *co-processors* [12] (see Figure 1 for a simplified representation of the two integration paradigms).

The European Commission has identified HPC centres as natural deployment sites for these quantum machines [13][14], as they provide a suitable infrastructure to empower such integration. The combination of QCs with HPC technologies is in fact believed to be the key to unlocking *useful quantum computation*. HPC centres also offer a controlled environment, which is still necessary to preserve the fragile quantum states from interactions with the outside world. None of the computing platforms under development has in fact demonstrated operational performance that is resilient to ambient conditions (either in terms of temperature, pressure, vibrations, or presence of electromagnetic fields), and the idea of deploying QCs on edge nodes, or mobile devices in uncontrolled environments, albeit suggestive, remains highly unpractical.
Regardless of the NISQ integration paradigm, even when fault-tolerance will be available, QCs will likely not replace traditional HPC resources, but they will rather cooperate with them to meet operational demands, possibly in a heterogeneous environment with CPUs, GPUs, and specialized hardware. It is therefore critical for organizations to prepare themselves with the necessary tools, skills, and capabilities that span both the HPC and quantum computing domains, in order to fully exploit the benefits that QCs will offer in the years ahead. HPC centres will serve as hosting sites for future QCs, but they already represent a strategic asset for conducting impactful R&D activities in the field. Firstly, the HPC resources enable emulating large-scale quantum computers that are still expensive, time-limited, and difficult to access, providing opportunities for developing in-house expertise, ideas, and insights that would be unattainable otherwise. Secondly, HPC centres currently offer high-end digital computing resources that can be utilized to test hybrid quantum-classical approaches for real-world problems, even within a loose-integration framework.

Regardless of the NISQ integration paradigm, even when fault-tolerance will be available, QCs will likely not replace traditional HPC resources, but they will rather cooperate with them to meet operational demands, possibly in a heterogeneous environment with CPUs, GPUs, and specialized hardware. It is therefore critical for organizations to prepare themselves with the necessary tools, skills, and capabilities that span both the HPC and quantum computing domains, in order to fully exploit the benefits that QCs will offer in the years ahead.

A PRAGMATIC RESEARCH APPROACH

The Leonardo Labs aim to assessing the potential benefits of combining QC with HPC in business-related use cases. Our mission is to build a comprehensive and pragmatic understanding of the capabilities, trends, and limitations of the technology to provide informed and unbiased recommendations that could steer future strategic initiatives within our organization. The activities are conducted mostly internally and through national or international collaborations with strategic research partners.
We focus on two main research streams, both targeting quantum algorithms. The first stream deals with benchmarking algorithms and software tools that could enable assessments across a spectrum of NISQ hardware platforms. This stream is designed to build in-house skills and to prepare a workforce that is responsive to rapid developments in the field. In practical terms, this endeavour is being pursued through the implementation of lighthouse projects that identify use cases with industrial significance. The activities are carried out by leveraging the proprietary *davinci-1* HPC [15] and the most mature quantum platforms in terms of hardware and software stack, user ecosystem, accessibility, and usability. Another active research stream focuses on experimenting quantum-inspired algorithms, that take inspiration from quantum computing methods but can run solely on HPC hardware resources. Unlike the full-quantum stream, which features a more exploratory focus, the challenge here is to address problems related to real business needs of Leonardo, and to identify possible computational advantages over more traditional approaches. We have currently identified a range of case studies in areas such as *computer vision, quantum chemistry, logistics*, and *telecommunications* that belong to the following problem domains: *combinatorial optimization, simulation,* and *quantum Machine Learning* (ML). Hereby, a portfolio of use cases:

- Target detection from remote sensing images with quantum Support Vector Machine (SVM);
- Resource management problems via variational quantum algorithms;
- Feature extraction from images with quantum annealers;
- Computational fluid dynamics via linear algebra quantum solvers;
- Routing and Traffic optimization via quantum and quantum-inspired formulations;
- Quantum chemistry/materials science for internal needs via quantum/classical hybrid framework;
- Clustering with quantum ML algorithms.

To effectively advance towards real-world applications that can impact the value chain, it is however critical to identify further potential applications that feature industrial significance. To this end, we seek to promote collaboration and discussions with Leonardo production units that have interesting use cases.

LIGHTHOUSE PROJECTS

In the attempt to foster this collaboration, we present here results obtained from two selected small-scale, big-picture pilot projects.

LP1: Target Detection in Hyperspectral Images using a Quantum Annealer

This project focuses on the problem of detecting specific types of vegetation from the observation of terrain hyperspectral images obtained from remote sensing instruments. Given one or more images, provided as matrices of pixels, the aim is to generate a supervised classifier capable of identifying pixels of the target vegetation class with the best performance possible. As usual in remote sensing applications, a challenging working condition is given by the reduced set of labelled samples from which to learn. At the same time, the issue of learning from few data is an ideal niche for testing NISQ quantum computers.

The terrain target detection problem is a relevant problem for Leonardo that is typically addressed via classical state-of-the-art deep learning (DL) methods. Here, instead, we try to address it by making use of quantum machine-learning (ML) approaches based on support vector machine (SVM) algorithms.
In this case, the SVM is preferable to DL algorithms as it relies on the margin maximization principle that is less sensitive to overfitting, especially when few training data are available [16]. The questions we try to address in this work are the following:

- Is it possible to use quantum-inspired approaches to train an SVM?
- Can quantum SVMs provide advantages when compared to classical ones for our use case?

To answer such questions, we focus on studying the generalization power of SVM classifiers either trained via classical methods with known libraries, or via SVM classifiers obtained first by reformulating the standard training problem as a quantum-inspired Quadratic Unconstrained Binary Optimization (QUBO) problem, and then by solving this problem with a quantum annealer.

Methods & Computing Tools

The key idea behind SVMs is to find a hyperplane that best separates the data into different classes. In two-class problems, the hyperplane is chosen such that it maximizes the margin between the closest data points of each class, known as support vectors.
This results in robust and effective decision boundaries, even when the data is not linearly separable. SVMs are also capable of handling non-linear data using kernel functions. Training a SVM on a given training-set $D_{train}=\{(x_i,y_i): i = 0,\dots,N-1\}$ (with $x_i \in R^d$ and $y_i = \pm 1$ being the feature vector and its label), consists in finding the optimal real coefficients $\{\alpha_i\}$ of the following quadratic problem.

$$L = \frac{1}{2}\sum_{i,j}\alpha_i\alpha_j\, y_i y_j\, k(\boldsymbol{x}_i,\boldsymbol{x}_j) - \sum_i \alpha_i \qquad (1)$$

subject to

$$0 < \alpha_i < C; \quad \sum_i \alpha_i y_i = 0 \qquad (2)$$

with C being a regularization parameter, and $k(x_i, x_j) = e^{-\gamma \|x_i - x_j\|^2}$ the kernel function chosen in this work, with γ a hyperparameter. The prediction for an arbitrary feature sample x can then be made by evaluating the decision function

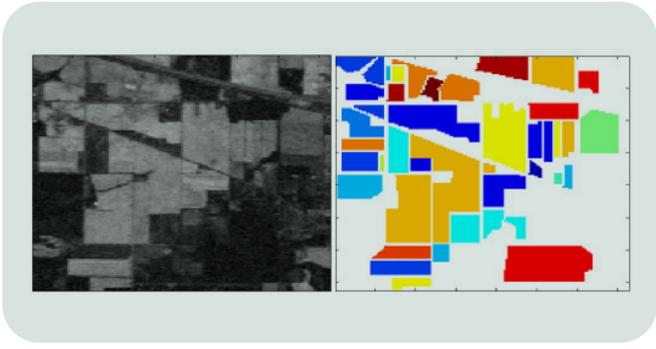
$$f(\boldsymbol{x}) = \sum_i \alpha_i\, y_i\, k(\boldsymbol{x}_i,\boldsymbol{x}) + b \qquad (3)$$

where the bias $b := b(\alpha_i, x_i, y_i)$ is taken as function of the optimal coefficients $\{\alpha_i\}$ and training data $\{(x_i, y_i)\}$ [16]. The sign of f(x) predicts the class label of x.

The QUBO problem corresponds to minimizing the cost function

$$H = \sum_{i\le j} w_i\, Q_{ij}\, w_j \qquad (4)$$

with $w_i \in \{0,1\}$ the binary variables of the optimization problem, and Q the QUBO upper triangular real-valued matrix. It is possible to reformulate the standard optimization problem in (1) into a QUBO problem which can then be used to train the SVM with a quantum annealer [16]. We refer to the SVM models based on QUBO as q-SVM, and to those based on classical approaches as c-SVM. With the appropriate manipulations and change of variables, the minimization problem (1) under conditions (2) can be reformulated in an appropriate QUBO problem (4).
The quantum solver used here for the optimization of the q-SVM is the D-Wave Advantage QA with 5000+ qubits with an average connectivity of 15 [17]. Details on the implementation will be discussed in a separate technical paper [18]. For the optimization of the c-SVM we used the scikit-learn python package which relies internally on the LIBSVM library [19].



2 - Sample band of the Indian Pines dataset (Left panel) and corresponding ground-truth (Right panel)

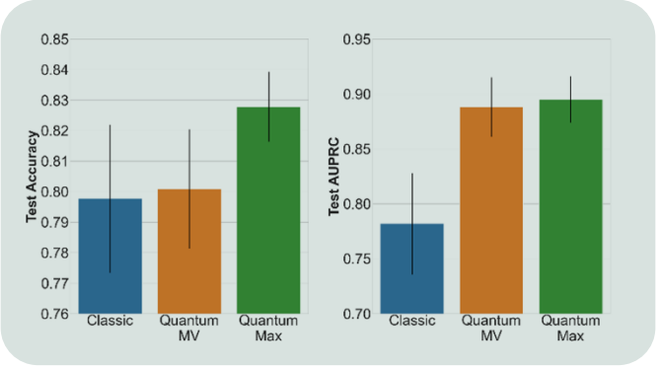
Dataset

The publicly available Indian Pines dataset [20] is used to train the SVM models. This dataset is chosen due to its popularity in the ML community, but we could seamlessly extend our work to deal with custom datasets for specific needs. The dataset comprises a single landscape in Indiana (USA) covering roughly 2×2 miles, with a sampling resolution of 145×145 pixels, obtained through an AVIRIS hyperspectral sensor. Each pixel exhibits 200 spectral reflectance bands corresponding to different portions of the electromagnetic spectrum.

The dataset encompasses a range of vegetation coverings labelled as 16 distinct classes, plus a background class, as from the ground truth reported in Figure 2. In our study, we focus on the corn class, utilizing a training set made of 50 pixels in a balanced configuration (25 corn, 25 no-corn classes).
This choice is motivated by the objective of testing the SVM method under the constraint of few training data, as well as by the current capabilities of the quantum annealer. Our test set contains 2000 (balanced) pixels. Given that the dataset's pixels (from any class) belong to the same area of a single image and they are therefore quite correlated, we did not prepare any validation set.

Results & Discussion

The classification pipeline includes three different phases: calibration, training, and testing. In the calibration phase we select the optimal hyperparameters for the SVMs by performing a grid search to maximize the score function defined below on the training set. To feed the SVM models we employ few principal components of the original hyperspectral spectrum associated to each pixel. The training step is then performed using the optimal hyperparameters obtained from calibration.
To provide a statistical dimension to our analysis, we generate *c*-SVM and *q*-SVM models starting from 100 different training sets D_{train} prepared with the recipe described in the previous section by randomly picking from a pool of pixels not used in the test set. It is worth noting however that the quantum annealer can be natively employed as a statistical sampler of suboptimal models.
Unlike the single optimal solution provided by the scikit-learn solver for the *c*-SVM case, a range of suboptimal models was therefore generated in the *q*-SVM case for each training set. This is shown for example in Figure 3, where we compare the aggregated score function values of the best suboptimal q-SVM models (maximum-based) generated by the quantum annealer for the different sets D_{train} against the *c*-SVM counterparts.
The results suggest that if we were able to find a proper strategy to extract such models directly from the training set, we would have a quantum-annealing-based recipe for making models with superior generalization performance compared to those obtained by standard state-of-the-art optimization procedures when the training data are limited. Various strategies are now under investigation. A simple majority-voting-based (MV) approach already allows us to beat classical results for all score functions, with marked improvements on the AUPRC metric [21].
These findings are highly encouraging as they indicate clear potential benefits in employing a quantum annealer for the present application.
Future work will be conducted to assess this method on selected Leonardo datasets, integrating this approach into the *HyperHound* artificial intelligence framework for target detection of hyperspectral images already developed at the Leonardo Labs [22].



3 - Mean and standard deviation of the test accuracy and AUPRC for the c-SVM and q-SVM models aggregated over 100 seeds. The quantum MV and Max labels indicate majority-voting-based and maximum-based q-SVM models

LP2: Power Management of Antenna Networks with Universal Gate-Based Quantum Computers

We devise here a combinatorial problem in the field of telecommunications that consists in the power management of a network of omnidirectional antennas covering a geographic area of interest.
The service provider should guarantee maximum coverage at high power, while minimizing overlap regions covered by multiple signals at high power. We indeed assume that operating all the antennas at the maximum power corresponds to an unwanted waste of energy with negative economic, strategic, or environmental implications.
This model can be applied for example to emergency scenarios where severe limitations on the available resources occur. For simplicity, we consider antennas with only two possible power levels, referred to as *low* and *high*. To model the two competing requirements, we rely on the following Ising form [23]:

$$H_S = \frac{1}{2}\sum_{i,j=1}^{N} J_{ij}\, s_i s_j - \xi \sum_{i=1}^{N} B_i s_i \qquad (5)$$

where $s_i \in \{-1,+1\}$ represent the power level on each antenna, and $\xi \in R^+$ is a regularization term that modulates the relative strength of the two terms in (5). The terms J_{ij} and B_i correspond to the pairwise coupling strength and the local effective field of the Ising model respectively. Both terms are defined a priori and problem dependent. For the application under investigation, the matrix elements are defined as $J_{ij} \propto B_i \cap B_j$, with B_i being the signal coverage area (here taken as circular) of site *i*. Optimal solutions of the original problem are encoded in the bitstring $S = \{s_1, s_2 \ldots s_N\}$ which minimizes H_S.
If operating at the maximum power everywhere is reputed more important than reducing power consumption, we set $\xi > 1$, otherwise we set $\xi \leq 1$.
Real-world applications can feature a more complicated structure, e.g. antennas can be directional and can operate over different frequencies (5G protocol). However, this minimal example already contains all important ingredients to illustrate how to treat this kind of problems with quantum algorithms.

Methods: the QAOA algorithm

To tackle this combinatorial optimization problem, we use the popular *quantum approximate optimization ansatz* (QAOA), introduced in [24]. The QAOA has been formulated within the universal gate-based framework for quantum computation and is considered a promising candidate to achieve quantum advantage [25].
It is a noise-resistant hybrid algorithm that uses a classical optimizer to prepare a quantum state that approximates the ground state of a given problem. In the QAOA, we introduce a quantum operator $\hat{H}_S$ obtained by H_S promoting the binary variables s_i to quantum mechanical operators. Given a quantum state $|\beta,\gamma\rangle$ parametrized in terms of continuous variable vectors β and γ, we introduce a cost function given by the expectation value of $\hat{H}_S$

$$f(\boldsymbol{\beta},\boldsymbol{\gamma}) = \langle \boldsymbol{\beta},\boldsymbol{\gamma} | \hat{H}_S | \boldsymbol{\beta},\boldsymbol{\gamma} \rangle \qquad (6)$$

Upon a suitable choice of procedure to generate the quantum state $|\beta,\gamma\rangle$, we can approximate the ground state energy E_GS by minimizing $f(\beta,\gamma)$, knowing that $f(\beta,\gamma) \geq E_{GS}$. The state resulting from the minimization is then completely determined by $(\beta^*,\gamma^*) = \arg\min_{\beta,\gamma} \langle \beta,\gamma | \hat{H}_S | \beta,\gamma \rangle$. The calculation of the expectation value is the operation that can be performed on a quantum computer, while the minimization procedure in the parameter space is performed using classical algorithms. The state is practically generated by acting on the superposition state $|+\rangle^{\otimes n} = \left(\frac{|0\rangle + |1\rangle}{\sqrt{2}}\right)^{\otimes n}$ with parametrized unitary gates as

$$|\boldsymbol{\beta},\boldsymbol{\gamma}\rangle = U(\boldsymbol{\beta},\boldsymbol{\gamma})\,|+\rangle^{\otimes n} \qquad (7)$$

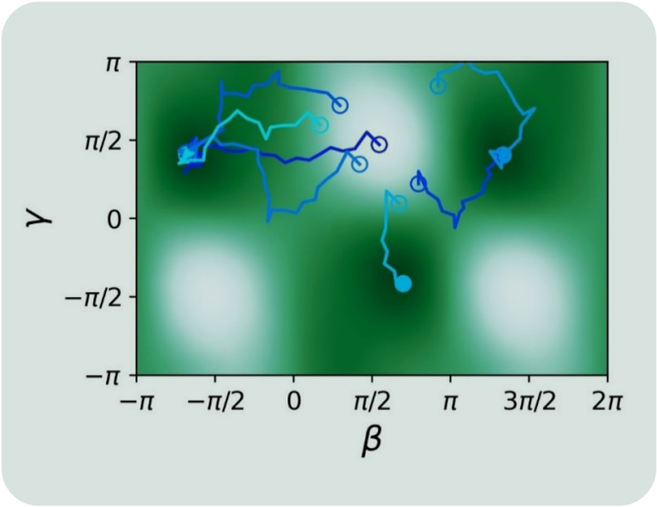
The explicit expression for the unitary gate is

$$U(\boldsymbol{\beta},\boldsymbol{\gamma}) = \prod_{n=1}^{p} e^{-i\beta_n \sigma_x} e^{-i\gamma_n H_S} \qquad (8)$$

where p indicates the depth of the circuit, σ_x is a Pauli matrix, and the optimization occurs in the 2p-dimensional space of parameters (β,γ). This ansatz has a layered structure reminiscent of artificial neural networks. At each optimization step, all the qubits of the state $|\beta,\gamma\rangle$ are measured a suitable number of times. Each measurement returns a bit string whose energy can be computed straightforwardly using (5). Averaging over the outcomes of the measurements, we get an estimate of $f(\beta,\gamma)$.

Computational tools: HPC for QAOA

Quantum approximate optimization algorithms are designed to leverage both the current NISQ hardware and the HPC resources through integrated hybrid workflows. High-Performance Computing are in fact vital to support the classical exploration of the high-dimensional space of parameters, speeding up the search of the global minimum of the non-convex function $f(\beta,\gamma)$. As the number of layers p increases, finding such global minimum becomes more and more challenging as the optimization runs typically get stuck into local minima. To increase the chances of finding the global minimum, efficient parallelization strategies are required. Along this line of thinking we develop a framework to initialize a “swarm” of independent walkers with different starting points that can run in a highly-parallel fashion through an Message Passing Interface (MPI) protocol [26].
After the execution of the parallel framework, we select the walker that exhibits the lowest value of the loss function. HPC systems allow to run the QAOA algorithms on emulated QCs.
Although the emulation of quantum circuits on a classical computer is limited by the exponential memory footprint required by the quantum states, the use of multiple, high-end GPUs available in HPC clusters enables the emulation of states with 20+ qubits in a reasonable amount of time.

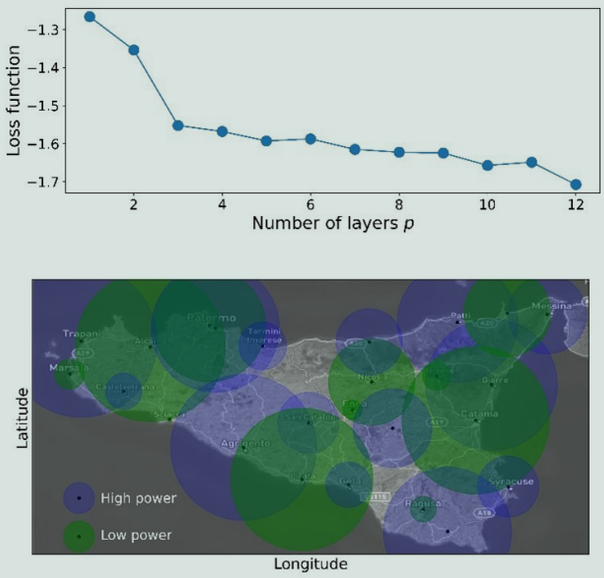


4 - Paths of parallel walkers exploring the loss function landscape for the antenna problem with p=1, in the classical minimization procedure. The empty/filled markers indicate the starting/final points along each path

Results & Discussion

We devise an artificial yet prototypical problem instance where 25 antennas are located in proximity of the largest cities in Sicily.
The radius of action of each antenna is generated randomly within a proper range of values, while ξ is set to 0.25. We tackle this problem by emulating a gate-based QC through the state vector Aer emulator implemented in the *qiskit* package [27] running on the Leonardo HPC cluster *davinci-1* [14]. In future work we will also make use of real QCs.
In Figure 4, we show the optimization paths of seven walkers in the *p = 1* case as to illustrate how to exploit the parallel framework introduced above. For each walker, an MPI process is created that uses multiple CPUs and one GPU. Each walker typically takes 25-30 minutes to perform 100 steps of the COBYLA optimizer [28].
At the production stage, we use a swarm of 32-walkers initialized with a properly distributed set of parameters in the 2p-dimensional space.

In the upper panel of Figure 5, we show the behaviour of the loss function versus the layer number up to *p = 12*. As expected, the loss function decreases with the number of layers thus suggesting a strategy for systematic improvement of the QAOA solutions.
From the qualitative point of view our results confirm those reported in previous studies [29]. In the bottom panel of Figure 5, we show the best numerical solution obtained from our calculations in graphical form. Interestingly, our solution well approximates the ground state of the problem instance at hand, with an approximation ratio of 98% in terms of the cost function. We will conduct future works to investigate how to scale up the size of the problem instances that can be treated with this approach while preserving the solution quality here reported.



5 - Convergence of the QAOA loss function as a function of the number of layers p (top panel). Solution of the problem with 25 antennas located in correspondence of the largest cities in Sicily (bottom panel)

CONCLUSIONS

In this work we have presented insights into the quantum computing initiatives conducted at the Leonardo Labs, by emphasizing the importance of combining HPC and QC resources to enable utilization of quantum computing hardware and boosting its developments. It is in fact critical to preside over this technological area, as practical realizations of quantum advantage are expected to emerge in the framework of hybrid quantum-classical approaches. The activities of the Lab are hence shaped to maximally exploit this scenario and are carried out both (1) to prepare the organization to exploit QCs for in-house applications promptly and effectively when they become profitable, and (2) to foster the development of unconventional quantum-inspired methods deployed on proprietary HPC to solve pressing near-term internal challenges.

We hence present two lighthouse projects relevant to Leonardo. First, we explore the use of a quantum annealer as a trainer of suboptimal SVM models to generate robust supervised classifiers of vegetation types from hyperspectral remote sensing images.
Our results indicate that these classifiers can outperform classical SVM models trained using state-of-the-art packages in terms of generalization capacity when training data is limited. Second, we investigate the optimization of the power configuration of an antenna network. We focus on a simplified problem formulation that has potential for generalization to complex real-world applications in the telecommunications field.
We evaluate the capability of QAOA algorithms to solve such problems when recast into QUBO forms as the size of the problem increases, leveraging emulation on Leonardo cluster *davinci-1*. We find the in-house MPI-based parallel workflow to be critical for extracting good approximate solutions.
This study establishes the groundwork for future developments that could enable the solution of such problems at industrial scale via quantum-inspired methods on HPC.
We hope this work could stimulate internal and external collaborations to challenge this new computing paradigm on internal use cases.

ACKNOWLEDGMENTS

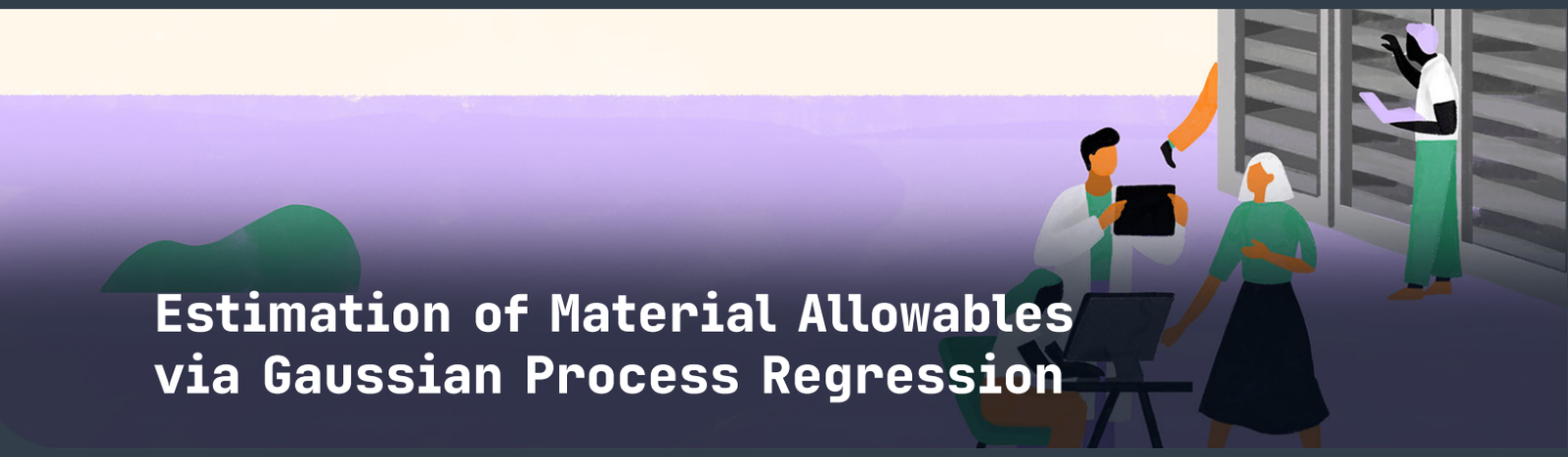
We thank Dr. Alessandro Garibbo and Dr. Roberto Agrone at Leonardo for engaging in discussions related to identifying antenna management scenarios.
We also thank CINECA for supporting the activities of the target detection project.

Daniele Dragoni: daniele.dragoni.ext@leonardo.com

REFERENCES

[1] J. Preskill, “Quantum computing in the NISQ era and beyond”, Quantum 2, 79 (2018)
[2] The quantumzoo website. [Online]. Available: https://quantumalgorithmzoo.org/
[3] P.W. Shor, “Algorithms for quantum computation: discrete logarithms and factoring”, IEEE Comput. Soc. Press, 124 (1994)
[4] D. Beckman et al., “Efficient networks for quantum factoring”, Phys. Rev. A, 54, 1034 (1996)
[5] The IBM webpage on 00 Qubit-Plus Quantum Processor. [Online]. Available: https://newsroom.ibm.com/2022-11-09-IBM-Unveils-400-Qubit-Plus-Quantum-Processor-and-Next-Generation-IBM-Quantum-System-Two
[6] The database webpage on Quantum Computer Roadmaps. [Online]. Available: https://databaseline.tech/quantum.html#references
[7] Google Quantum AI, ”Suppressing quantum errors by scaling a surface code logical qubit”, Nature 614, 676 (2023)
[8] J.F. Bobier et al., “What Happens When ‘If’ Turns to ‘When’ in Quantum Computing?”, BCG (2021). [Online]. Available: https://www.bcg.com/publications/2021/building-quantum-advantage
[9] F. Arute et al., “Quantum supremacy using a programmable superconducting processor”, Nature 574, 505 (2019)
[10] H.S. Zhong et al., “Phase-Programmable Gaussian Boson Sampling Using Stimulated Squeezed Light”, Phys. Rev. Lett. 127, 180502 (2021); Y. Wu et al., “Strong Quantum Computational Advantage Using a Superconducting Quantum Processor”, Phys. Rev. Lett. 127 180501 (2021)

[11] L.S. Madsen et al., “Quantum computational advantage with a programmable photonic processor”, Nature 606, 75 (2022)
[12] T.S. Humble et al. “Quantum computers for high-performance computing”, IEEE Micro 41, 15 (2021)
[13] EuroHPC JU webpage on Selection of six sites to host the first European quantum computers. [Online]. Available: https://eurohpc-ju.europa.eu/selection-six-sites-host-first-european-quantum-computers-2022-10-04_en; The HPCQS website. [Online]. Available: https://www.hpcqs.eu/
[14] V. Bartsch et al., “Quantum for HPC – The impact of Quantum Computers on HPC applications and the integration of quantum computers in HPC centres”, ETP4HPC white paper (2021)
[15] The Leonardo webpage on “the HPC system davinci-1”. [Online]. Available: https://www.leonardo.com/en/innovation-technology/davinci-1
[16] G. Cavallaro et al., “Approaching Remote Sensing Image Classification with Ensembles of Support Vector Machines on the D-Wave Quantum Annealer”, IEEE IGARSS 2020-2020, 1973 (2020)
[17] Dwave webpage on Advantage Processor Overview. [Online]. Available: https://www.dwavesys.com/media/3xvdipcn/14-1058a-a_advantage_processor_overview.pdf
[18] “Target Detection in Hyperspectral Images using a Quantum Approaches”, In preparation (2023)
[19] F. Pedregosa et al., “Scikit-learn: Machine Learning in Python”, JMLR 12, 2825 (2011)
[20] M. F. Baumgardner et al., “220 Band AVIRIS Hyperspectral Image Data Set: June 12, 1992 Indian Pine Test Site 3”, Purdue University Research Repository, 10, 991 (2015)
[21] J. Jesse et al., “The relationship between Precision-Recall and ROC curves”, Proceedings of the 23rd international conference on Machine learning, 233 (2006)
[22] R. Di Carlo et al., “HyperHound: a Framework for Hyperspectral Image Analysis and Target Detection using Deep Learning Models”, POLARIS Innovation Journal - Issue 48, 2023
[23] Note that Ising and QUBO forms are connected by a simple variable substitution
[24] E. Fahri et al., “A Quantum Approximate Optimization Algorithm”, arXiv:1411.4028 (2014)
[25] L. Zhou et al., “Quantum Approximate Optimization Algorithm: Performance, Mechanism, and Implementation on Near-Term Devices”, Phys. Rev. X 10, 021067 (2020)
[26] L. Dalcin et al., “mpi4py: Status Update After 12 Years of Development”, Comput. Sci. Eng. 23, 47 (2021)
[27] Qiskit: Open-Source Quantum Development. [Online] Available: https://qiskit.org/
[28] M. J. D. Powell. “A view of algorithms for optimization without derivatives”, Cambridge Uni. Tech. Rep. DAMTP (2007)
[29] Y. Chai et al., “Shortcuts to Quantum Approximate Optimization Algorithm”, Phys. Rev. A 105, 042415 (2022)



Estimation of Material Allowables via Gaussian Process Regression

Roberta Cumbo[1], Antonio Baroni[2], Alfredo Ricciardi[2], Alessandro Nicolosi[3], Abhishek Kumar[1], Stefano Giuseppe Corvaglia2

[1]Leonardo Labs - Material Technologies, [2]Leonardo - Aerostructures Division,
[3]Leonardo Labs – Applied Artificial Intelligence

The accuracy in the definition of material allowables is highly important in order to guarantee the integrity of the structure. Thus a high number of experimental tests are required in order to consider most of the uncertainties associated to the entire manufacturing process. Several methodologies have been explored in literature by combining simulation-based solutions with a reduced set of test data but all the proposed alternatives cannot yet find any clear applicability in the industrial field. The solution proposed hereby aims to build a Machine Learning framework based on the knowledge of analytical principles of Fracture Mechanics of composites, trained by a minimum set of physical tests. The presented approach can support the design and characterization phase with a faster tool and reduced impact on the overall costs.

INTRODUCTION

The mechanical characterization of composite materials is at the bottom of the building-block pyramid, which is the standard approach used for qualification and certification of composite structures. The certification is established by the guidelines described in the Composite Military Handbook 17 [1] (CMH-17).
Coupon specimens are tested in compliance with the international standards from the American Society for Testing and Materials (ASTM), which normally require at least 5 samples for each test case in order to take into account possible source of uncertainties during the test execution.
In addition, other factors have high influence on the estimation of the ultimate strength of such specimens, such as the uncertainties associated to the constituent's properties and inaccuracies during the manufacturing process. For these reasons, the CMH-17 requires a minimum number of 18 specimens to be tested for each stacking-sequence and a given test case in order to guarantee safe estimation of the ultimate strength for the integrity of the full structure. The statistical value of ultimate strength defines the material allowables.
The whole process of manufacturing and testing leads to delays in the certification and in the design of a new structure. Nowadays, engineers approach some preliminary steps by using advanced simulation with Finite Element Analysis (FEA) codes. However, the "virtual" field of composite materials is computationally expensive and can require several physical parameters (which are not always available), as much as several trial-and-error iterations in order to reach high correlation with the real behaviour.
The world of simulation has been showing improvements in the last years by providing more features that are closer to the real behaviour of composite materials.

Recently, data-driven approaches have been also explored in order to further accelerate the design phase. Particular attention was given to Machine Learning (ML) algorithms. Furtado et al. [5] proposed the comparison of XGBoost, Random Forest (RF), Artificial Neural Network (ANN) and Gaussian Process Regression (GPR) for the prediction of the Open Hole Tension (OHT) specimens on virtual data, by reaching very low regression error for a small number of data samples.

The usage of GPR is particularly suitable in the field of composite materials because of the statistical nature of the algorithm, which can indeed consider the uncertainties associated to bulk materials and manufacturing errors. The choice of input features presented in [5] is established by the knowledge of analytical Fracture Mechanics principles. In [6], the approach for compression test cases is extended by linking the approach to the analytical solution for open-hole compression strength which considers fibres micro-buckling failure.

The work brings also some observations related to the selection of an optimal training scenario, i.e. subset of laminates, to be physically tested in order to train a non-linear regression algorithm for the prediction of other unknown laminates with less uncertainty.

This contribution aims to extend the previous work by proposing an optimal training scenario that can be generalized for different unidirectional materials. The number of physical test data is thus fixed and covers the full lamination plane by reaching the optimal ratio between the number of samples and the regression error. A set of three materials are considered for un-notched and open-hole tension. The experimental data at lamina level are used to calibrate the model in a FEA commercial software for the creation of a simulated database of different stacking sequences. The paper is structured as follows: a brief summary about GPR and analytical model for OHT is presented.

Then, the simulated Design of Experiments (DOEs) in terms of materials and stacking sequences is described. Finally, the results for a fixed number of training samples are shown and compared to the ones obtained in [6].

Gaussian Processes for non-linear regression applied to Machine Learning

Given a dataset of s samples, a linear regression can be expressed as:

$$\boldsymbol{y}^{(i)} = \boldsymbol{x}^{(i)}\theta + \boldsymbol{\epsilon}^{(i)} = \boldsymbol{f}^{(i)}(\boldsymbol{\theta}, \boldsymbol{x}) + \boldsymbol{\epsilon}^{(i)}, \qquad i = 1{:}s \qquad (1)$$

Where $\theta \in R^{(p+1)}$ is the vector of weights, $x \in R^{s \times (p+1)}$ is the matrix of features and ϵ represents additive noise which is usually Gaussian. The non-linear formulation of (1) is obtained by applying a projection into a high-dimensional feature space[7]. Under the assumption of weights and bias as mutually independent Gaussian distribution, the Bayesian linear regression (BLR) of (1) converges to a Gaussian Process (GP):

$$f \sim GP(m(\boldsymbol{x}), k(\boldsymbol{x})) \qquad (2)$$

with m, k as mean and covariance function respectively. In the framework of supervised learning, a Bayesian regression approach can be formulated as follows [7]: the GP is used as *prior*, which is updated with training samples x_{train} resulting in the posterior, i.e. prediction of $f(x^*)$ corresponding to the unobserved samples data x^*. The GP *posterior* can be summarized with the following set of equations:

$$\begin{cases} m_p(x) = m(x) + \Sigma(x_{train}, x)^T \Sigma(x_{train}, x)^{-1} \left(f(x_{train}) - m(x_{train})\right) \\ k_p(x, x') = k(x, x') - \Sigma(x_{train}, x)^T \Sigma(x_{train}, x)^{-1} \Sigma(x_{train}, x') \end{cases} \qquad (3)$$

From (3), the prior mean and covariance function are updated with training data and the uncertainty on the estimation identified by $k_p(x,x')$ is always less than the *prior* covariance $k(x,x')$. Normally, the estimation of functions close to the training data returns a small covariance, which is instead very high for samples located in the space of features out of the training domain. A comparison between the Bayesian Linear Regression and the Gaussian Process Regression (GPR) for composite materials can be found in [8]. One of the main advantages of the GPR is the high level of accuracy reached even for a small dataset.

Finite Fracture Mechanics principles for the prediction of ultimate Open Hole Tension strength

The failure of a composite laminate is determined by two main factors: the brittleness of the material and the ratio between the crack (if any) and the characteristic size of the tested specimen. When dealing with extreme cases of null or large crack, a stress or energetic criterion is considered, respectively. For intermediate cases, both those failure criteria should be considered.

Cornetti et al. [9][10] proposed a Finite Fracture Mechanics (FFM) criterion by coupling stress and energetic criteria for notched laminated in tension:

$$\begin{cases} \frac{1}{l}\int_{R}^{R+l} \sigma_{xx}(0,y)dy = \sigma_{UN} \\ \frac{1}{l}\int_{R}^{R+l} K_I^2(a)da = K_{IC}^2 \end{cases} \quad (4)$$

Where l is the extension of the crack, i.e. process zone, *R* is the characteristic length of the initial crack (equal to the radius of the hole for OHT specimens), σ_{UN} is the ultimate strength of the same specimen in un-notched conditions, K_I is the stress intensity factor and K_{IC} is the fracture toughness of the material. A schematic representation of notched specimen after failure is shown in Figure 1, where the process zone is clearly visible.

The formulation in (4) is considered in [6] in order to retrieve notched strength starting from the knowledge of un-notched values. The notched strength of an open hole coupon in tension for a fixed material results thus to be a function of the fracture toughness, un-notched ultimate strength and layup of the laminate:

$$\sigma_{OHT} = f(\sigma_{UNT}, K_{IC}, (\zeta_1, \zeta_2)) \quad (5)$$

The couple (ζ_1, ζ_2) indicates the lamination parameters, defined in the following section. Some of the assumptions behind (4) are valid for coupons with large ratio between width W of the specimen and diameter D of the hole. In this paper, coupons with sizes in agreement with ASTM D5766 are considered which fits the assumptions of the analytical framework.

1-Open Hole Tension specimen after failure

PREDICTIVE GPR FOR THE ESTIMATION OF MATERIAL ALLOWABLES

Design of Experiments

The dataset for the current study has been generated via Finite Element analysis through the commercial software Digimat [11]. The dataset is defined as follows:

- Test type according with ASTM D3039 [12] for Un-Notched Tension (UNT) and ASTM D5766 [13] for OHT;
- L stacking sequences according to manufacturing constraints for M materials.

For in-plane loads, the stacking sequences can be represented by two lamination parameters ζ_1, ζ_2 [14]:

$$\begin{cases} \zeta_1 = \frac{1}{h}\sum_{i=1}^{N} cos\,(2\theta_i) \\ \zeta_2 = \frac{1}{h}\sum_{i=1}^{N} cos\,(4\theta_i) \end{cases} \quad (6)$$

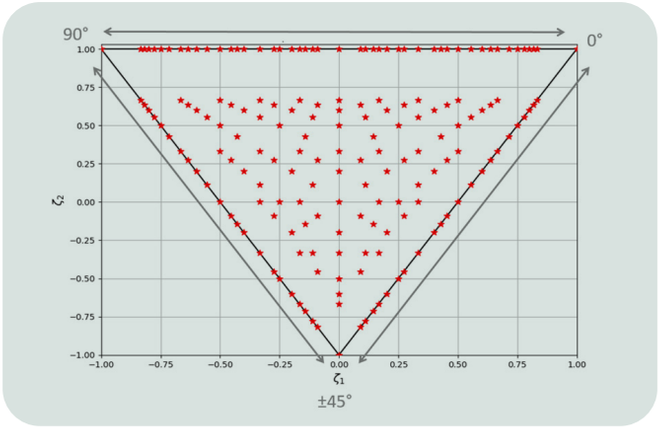
where h is the thickness of the laminate, *N* is the number of plies and θ_i is the orientation of i-th ply. As in [6], the following constraints are identified in order to reduce the range of existing stacking sequences within the range of the most commonly used laminates in industry:

- Number of plies ranging between 12 and 24;
- Balanced and symmetric laminates with ply orientation ±45°, 0°, 90°.

The total number of stacking sequences *L* within these constraints is 179. The distribution of the samples on the lamination plane is shown in Figure 2. The corners of the plane indicate the maximum percentage of plies for each orientation. When moving between corner e.g. (0,-1) to (1,1), the ratio between percentage of plies at 0° and percentage of plies at ±45° increases until reaching unidirectional laminate with 100% plies at 0° at the upper-right corner. Percentage values of some characteristic samples are listed in Table 1. The type of material investigated is fibre reinforced unidirectional tape. In particular, the following materials are considered:

- Hexcel 8552 IM7 [15];
- NCT4708 MR60H [16];
- S2/SP381 [17].

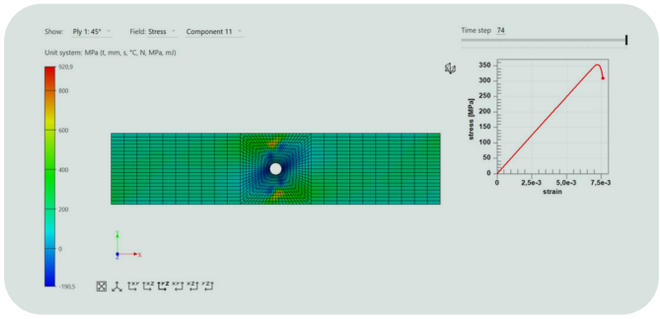
A database of 1074 (179 laminates x 2 test cases x 3 materials) simulated data has been created. The laminates are modelled via Digimat in agreement with ASTM. The Progressive Failure Analysis (PFA) [18] is considered for the evaluation of ultimate strength.
As illustrative example, the stress distribution of a general unidirectional laminate in OHT conditions is shown in Figure 3. For each sample, 18 simulations have been performed by varying the physical parameters of the material with a normal distribution and uncertainty level of 6% with respect to the baseline values.



2 - Lamination space (ζ_1, ζ_2). Samples: 179 laminates

(ζ_1, ζ_2)	0° %	±45° %
(1.0, 1.0) – *unidirectional*	100	0
(0.0, -1.0)	0	100
(0.0, 0.0) – *quasi-iso*	25	50
(0.0, 1.0)	50	0
(0.5, 0.0)	50	50

Table 1 – Percentage of orientation of plies for a set of (ζ_1, ζ_2) points



3 - Stress distribution at ultimate strength for unidirectional laminate along x-axis for OHT test. First ply at 45° of quasi-iso laminate. Simulation performed with Digimat

Training scenario for non-linear regression

The prediction accuracy for a fixed set of minimum and sub-optimal training samples is investigated in this contribution. The training scenario, defined by a subset of samples on the lamination plane, identifies the reduced number of laminates to be physically tested in order to obtain an accurate estimation over the full plane. A random sampling has been presented in [6] and briefly described below.
However, the random sampling approach depends on the dataset and thus is material dependent. In this contribution, the equal-spaced sampling has been adopted, in order to have a fixed training scenario that is independent from the material type.

Random sampling

The optimization criterion presented in [6] has been chosen by training a GPR for each instance subsets and by evaluating the Mean Absolute Error (MAE) on a predefined test data, which has been chosen ad-hoc by covering the full space including the most significant laminates with mixed plies at 0°, ±45° and 90°. The size of the training instances has been stated as a percentage of the full set of data and it counts 16 samples.

Fixed sampling

In this paper, 13 equal-spaced samples in the plane ζ_1, ζ_2 are selected. These points are chosen independently from the material type and are equally distributed over the lamination plane as shown in Figure 4 and Figure 5. The quantity of samples, i.e. 13, has been selected by maximizing the ratio between the regression accuracy and the number of samples. The corresponding stacking sequences in terms of percentage of ply orientations are listed in Table 2.

(ζ_1, ζ_2)	0° %	±45° %
(0.0, -1.0)	0	100
(-1.0, 1.0)	0	0
(0.0, 1.0)	100	0
(1.0, 1.0)	50	0
(-0.5, 0.0)	0	50
(0.0, -0.5)	12.5	75
(-0.5, 0.5)	12.5	25
(0.0, 0.0)	25	50
(-0.5, 1.0)	25	0
(0.0, 0.5)	37.5	25
(0.5, 0.0)	50	50
(0.5, 0.5)	62.5	25
(0.5, 1.0)	75	0

Table 2 - Fixed samples

Estimation of material allowables: results and discussion

The results of the estimation in terms of allowables and relative percentage error with respect to simulated reference data are shown in Figure 4 and Figure 5 for the material Hexcel 8552 IM7, by comparing the performances of both the random and the fixed sampling. In this paper, algorithms integrated in scikit-learn [19] Python package are used. The allowables are evaluated by using the following formula:

$$\hat{y} = \bar{y} - k\sigma_y \qquad (7)$$

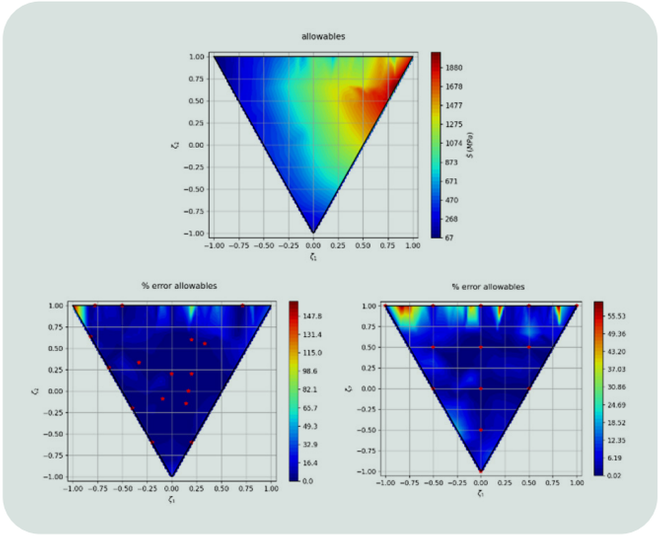
Where $\bar{y}$, σ_y are respectively the mean value and the standard deviation of uncertain data available for each laminate. *k* is a statistical parameter which depends on the data distribution. In this work, a normal distribution is adopted. The mean absolute error averaged over the full lamination plane is compared for both methods in Table 3. From Table 3, Figure 4 and Figure 5, it can be concluded that similar performances can be obtained with both methods and with a reduced number of training samples for the presented approach.
For the random sampling, the UNT results to be less accurate in terms of averaged MAE because only few samples cover the upper side of the plane, which is the one with highest uncertainty.
The prediction error obtained with fixed sampling for all the investigate materials is also shown in Table 4. Given the dependency of the OHT from the UNT data as in (5), the estimation in two ways:

- Standard: $\sigma_{OHT}=f(\sigma_{UNT}, K_{IC}, \textit{layup})$, where σ_{UNT} is simulated data;
- GP-UNT: $\sigma_{OHT}=f(\hat{\sigma}_{UNT}, K_{IC}, \textit{layup})$ where $\hat{\sigma}_{UNT}$ is estimated, i.e. Gaussian distribution $\hat{\sigma}_{UNT} \sim N(\bar{\sigma}_{UNT}, \Sigma_{UNT})$, with $\bar{\sigma}$ and Σ are respectively the mean and variance.

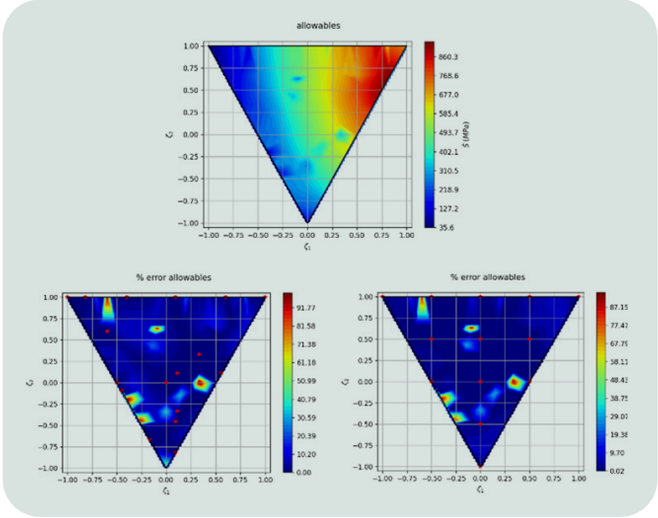
The second approach allows a further reduction of the number of data to be physically tested, since the estimation of OHT is performed based on estimated UNT properties. Only 13 samples of UNT are needed in this case. The results of the estimation for the materials NCT4708 MR60H are shown in Figure 6 as illustrative example.
The error maps for all the investigated materials show a value ranging between 15% and 20% for the centre of lamination space, which is a good result since the most of the laminates used in aerospace applications are located in that area. The areas close to the upper vertices are the ones with the highest uncertainty and this is in compliance with the real data since the laminates with 100% of plies oriented at 0° or 90° are usually the ones with more sources of uncertainties for the investigated tests.

	Random sampling [4]		Fixed sampling	
Test	n. training samples	MAE	n. training samples	MAE
UNT	16	77	13	57
OHT	16	23	13	17.4

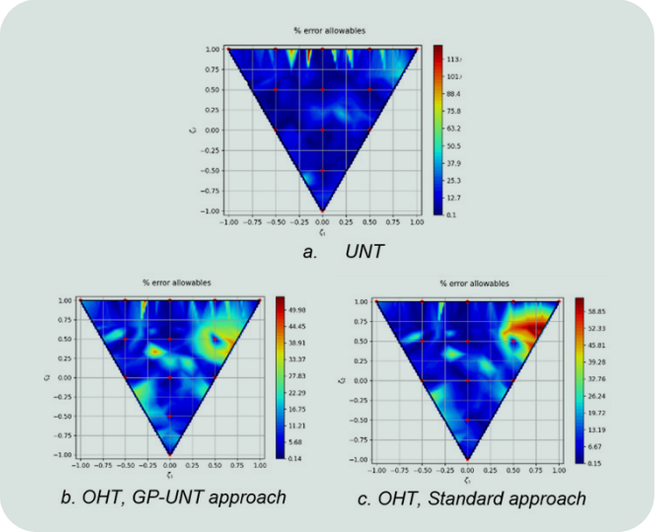
Table 3 - Estimation performances, material Hexcel 8552 IM7. Number of training samples and mean absolute error (MAE) for UNT and OHT



4 - UNT allowable values (top) and relative percentage errors with respect to simulated data (bottom). Results from random sampling (left) and fixed sampling (right)



5 – OHT allowable values (top) and relative percentage errors with respect to simulated data (bottom). Results from random sampling (left) and fixed sampling (right)



6 - Relative percentage error on the prediction, material NCT4708 MR60H

Material	Test	MAE	
Hexcel 8552 IM7	UNT	57.0	
		Standard	*GP-UNT*
	OHT	17.4	18.7
NCT4708 MR60H	UNT	97.0	
		Standard	*GP-UNT*
	OHT	58.7	50.5
S2/SP381	UNT	55.0	
		Standard	*GP-UNT*
	OHT	15.2	20.0

Table 4 – Averaged MAE of the estimation with fixed sampling for each material. Comparison between the two approaches for OHT

Moreover, the fixed sampling for the training set results to give good accuracy on the estimation of the three materials for the UNT test case. This means that the predicted values of the UNT strength can be re-used for the estimation of the OHT ultimate strength, allowing a further reduction of experimental data needed. In this case, ideally only 13 laminates under UNT conditions should be tested.

CONCLUSIONS

The application of GPR for the estimation of material allowables has been tested in this paper on simulated data for three fiber-reinforced composite materials. The aim is to improve the study on the usage of GPR for virtual material allowables by generalizing the approach independently from the material type. UNT and OHT test cases have been investigated. Contrarily to the previous work [6], the number of samples chosen as the training scenario is fixed and covers the full lamination plane by reaching the optimal ratio between the regression error and the number of samples. This approach has been validated on simulated data and it is estimated that it can reach 40% reduction of the number of physical tests with respect to standard industrial experimental campaign. Moreover, the knowledge of FFM principles allows to link the allowables of the notched specimen with the predicted values of un-notched laminates.

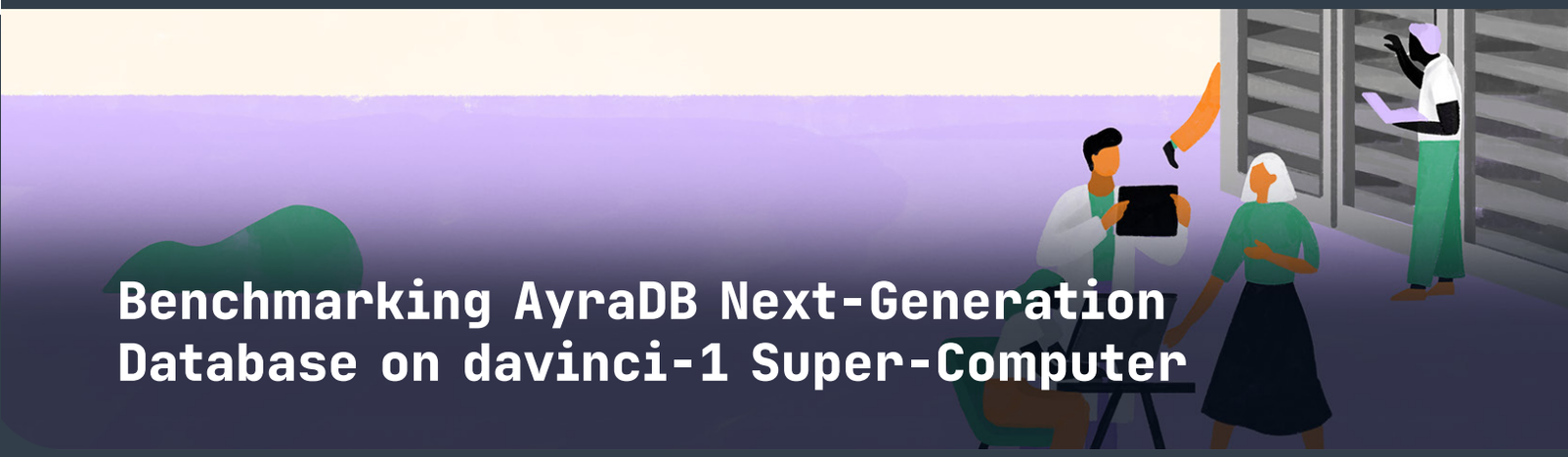
The approach allows a further reduction in the number of physical tests. However, the proposed approach needs to be validated on experimental data. which will be the scope of a future work. The described approach can be further refined by studying the uncertainty propagation of the estimation of un-notched allowables on the estimation of open hole strength.

Given the promising results shown in the paper, the direction of the industries and research centres operating in the field of design and certification should aim to cluster all the physical and numerical experiments collected during years of study with the aim to share the data with the composite material community. This would allow for more robust and more reliable applicability of the regression ML algorithms.

Roberta Cumbo: roberta.cumbo.ext@leonardo.com

REFERENCES

[1] Composites material handbook, Polymer Matrix Composites, Volume 1: Guidelines for characterization of structural materials. 1997.
[2] Cumbo, R., Baroni, A., Ricciardi, A., & Corvaglia, S. (2022). Design allowables of composite laminates: A review. Journal of Composite Materials, 56(23), 3617-3634.
[3] Van Vinckenroy G and De Wilde WP. The use of Monte Carlo techniques in statistical finite element methods for the determination of the structural behaviour of composite materials structural components. Compos Struct 1995; 32(1–4): 247–253.
[4] Arregui-Mena JD, Margetts L and Mummery PM. Practical application of the stochastic finite element method. Arch Comput Methods Eng 2016; 23(1): 171–190.
[5] Furtado, C., Pereira, L. F., Tavares, R. P., Salgado, M., Otero, F., Catalanotti, G., Arteiro, A., Bessa, M.A., Camanho, P. P. A methodology to generate design allowables of composite laminates using machine learning. International Journal of Solids and Structures, 2021, 233: 111095.
[6] Cumbo, R., Baroni, A., Ricciardi, A., Nicolosi, A., Corvaglia, S. Gaussian Process Regression for the prediction of material allowables, Proceedings of the 20th European Conference on Composite Materials, ECCM20, Vol. 4. June 2022, Lousanne, Switzerland.
[7] Rasmussen, C. E. Gaussian processes in machine learning. In: Summer school on machine learning. Springer, Berlin, Heidelberg, 2003. p. 63-71.
[8] Cumbo, R., Baroni, Kumar, A., Nicolosi, A., Corvaglia, S. Bayesian Machine Learning for faster design of composite structures. Proceedings of 73rd International Astronautical Congress (IAC). 2022.
[9] Cornetti, P., Pugno, N., Carpinteri, A., & Taylor, D. Finite fracture mechanics: a cou-pled stress and energy failure criterion. Engineering Fracture Mechanics, 73(14), 2021-2033 (2006).
[10] Camanho, P. P., Erçin, G. H., Catalanotti, G., Mahdi, S., & Linde, P. (2012). A finite fracture mechanics model for the prediction of the open-hole strength of composite laminates. Composites Part A: Applied Science and Manufacturing, 43(8), 1219-1225.
[11] Digimat Virtual Allowables, https://www.e-xstream.com/ (accessed 17 February 2023)
[12] Standard Test Method for Tensile Properties of Polymer Matrix Composite Materials, D3039/D3039M-08, ASTM International.
[13] Standard Test Method for Open-Hole Tensile Strength of Polymer Matrix Composite Laminates, D5766/D5766M-11 (Reapproved 2018), ASTM International.
[14] Tsai SW, Pagano NJ. Invariant properties of composite materials. Tech. Rep.; Air force materials lab Wright-Patterson AFB Ohio; 1968.
[15] E. Clarkson, Hexcel 8552 IM7 Unidirectional Prepreg 190 gsm & 35%RC Qualification Statistical Analysis Report, Report Number NCP-RP-2009-028 Rev B, April 2019.
[16] K.L. Poon, Newport Adhesive and Composite NCT4708 MR60H 300gsm 38%RC Unidirectional Qualification Material Property Data Report, NCAMP Test Report Number CAM-RP-2010-041 A, August 2011.
[17] J. Tomblin, J. McKenna, Y. Ng, K. S. Raju, B-Basis Design Allowables for Epoxy-Based Prepreg, 3M S-Glass Unitape S2/SP381, AGATE-WP3.3-033051-099, September 2001.
[18] Matzenmiller ALJTR, Lubliner J, Taylor RL, et al. A constitutive model for anisotropic damage in fiber composites. Mech Mater 1995; 20(2): 125–152.
[19] https://scikit-learn.org/ (accessed 22.07.2022)



Benchmarking AyraDB Next-Generation Database on davinci-1 Super-Computer

Roberto Morelli[1], Nicolò Magini[4], Carlo Cavazzoni[1], Chiara Francalanci[3], Paolo Giacomazzi[3], Paolo Ravanelli[2]

[1]Leonardo Labs - Applied Artificial Intelligence, [2]Cherrydata Srl,
[3]Politecnico di Milano - Dipartimento di Elettronica, Informazione e Bioingegneria, [4]Leonardo - Corporate

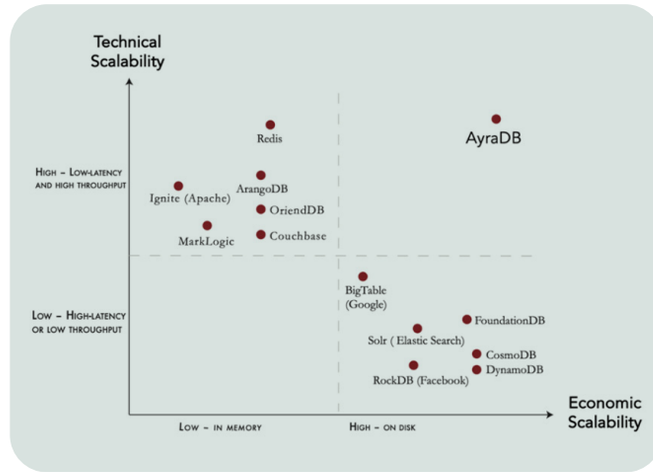
This paper describes a 1-year project that is aimed to test the performance and scalability of a new database called AyraDB on the davinci-1 facility. AyraDB is different from other databases, because it is fully peer-to-peer, which means there is no central node to coordinate other storage nodes, and it doesn't need to cache data in memory. This makes it much more scalable than other databases, like MongoDB and Redis, which feature central coordination nodes and/or need to cache data in memory. AyraDB is especially useful for storing and retrieving satellite data that can be very large and complex. The authors of this paper did perform tests on the davinci-1 high-performance computing (HPC) infrastructure that is designed for data-and processing-intensive applications. A set of 500 tests has been performed on servers with 8 cores, ranging from 1 to 20 servers, and with data sizes ranging from 10 GB to 500 GB. The results show that AyraDB can handle 1 million requests per second with just 13 servers, while its performance scales linearly as more servers are added. This beats the state-of-the-art by a factor of 5. Overall, the authors believe that AyraDB has the potential to enable new use cases in the space economy.

INTRODUCTION

This work discusses the need for new and more efficient data management technologies that support the exponential growth of big data projects [1].
A new database named AyraDB [2] is benchmarked to assess its performance and scalability.
AyraDB is unique as it is fully peer-to-peer, with no central node to coordinate other storage nodes, it has been developed by Cherrydata Srl [10] to enable the most demanding big-data use-cases. This technology targets the key requirements of horizontal scalability, the on-disk data storage with no caching, and the thrifty hardware consumption through greater performance.
Horizontal scalability means that AyraDB can easily accommodate growing capacity requirements by adding nodes to the database infrastructure. This allows the database itself to handle increasing amounts of data without any degradation in performance.
On-disk data storage with no caching means that AyraDB stores data directly on the disk, without keeping a copy in memory. This makes the database more efficient when dealing with large amounts of data, as it avoids the need for expensive and time-consuming caching operations.
Thrifty hardware consumption through greater performance means that AyraDB is designed to consume hardware resources more efficiently than other competing databases, while still providing high levels of performance. This allows organizations to reduce their hardware costs and energy consumption, while still meeting their data management needs.
All these features make AyraDB a good candidate for several applications. For example, satellite data are notoriously big data and raise several challenges from a different point of view concerning both analytic

applications (such as statistics or machine learning) and efficient and fast storing and data retrieval [3].
The goal of the research presented in this paper is to perform a large-scale test of AyraDB to verify and measure its performance and scalability edge. The tests are carried out on a davinci-1 high-performance-computing (HPC) facility that provides a perfect infrastructure to run AyraDB's tests. Indeed, davinci-1 is in-house managed, with total control of the infrastructure, thus enabling to design a test with no bottlenecks and providing dependable performance benchmarks. These tests are key to demonstrate the technical and business readiness of AyraDB, since the efficient management of large datasets represents a critical success factor.
Finally, these tests have helped to estimate the energy savings enabled by AyraDB. The results show that AyraDB can significantly reduce hardware and energy consumption. Also, in terms of long-term sustainability, the test can act as a proof of concept of the technical features of AyraDB and of its suitability for challenging applications. A number of application opportunities could open up, particularly in the space industry, enabled by the mix of performance, scalability, and favourable economics of AyraDB.

The survey of benchmarking efforts of big data systems can be found in [4]. According to the survey, scientific benchmarking studies mostly focus on a limited number of nodes (typically 3) and rarely test the maximum performance of databases or their scalability limits due to the high cost of large-scale tests. On davinci-1, the minor limitation on the server's usage favours extensive performance exploration. The tests provided in this work are focused on typical database operations like a key-value read/write load.
To benchmark the efficacy of a database solution, the number of these kinds of operations (throughput) is quantified by considering that larger number means better performance. To the best of our knowledge, the highest performance achieved by top-performing solutions is around 300k operations/s. Couchbase, for example, reaches this same level with a cluster of 10 servers, with minimal increase when doubling the number of servers [5].
MongoDB, on the other hand, can only reach 200k operations/s with 20 servers.
Scylladb [6] has also been benchmarked, with a maximum per-core throughput of 6.2k operations/core and an overall maximum throughput of 300k operations/s. However, this throughput is halved during write operations and is reduced by a factor of 3 with a load that limits the benefits of in-memory caching.
Finally, a benchmark for Riak database shows that its performance settles around 300k operations/s with an enterprise-significant number of keys in the database [7]. In terms of market analysis, the Hippj project conducted a thorough analysis and found that AyraDB is the only database that offers both technical and economic scalability, without any trade-offs between cost and performance or limitations on application scenarios. Figure 1 reports the market position of AyraDB in comparison with other competing database solutions. One of the major features of AryaDB regards the data storage that involves no trade-off between cost and performance and no limitations on the application scenarios.
The main feature of AyraDB is the data storage system which eliminates the need for caching in memory to achieve response times shorter than one millisecond. This high-performance capability can be achieved by using standard SSD hard drives, which makes it a cost-effective solution. By utilizing disk storage instead of memory, AyraDB reduces costs by a factor of 20.
Despite an extensive search, no comparable solutions have been found that can match AyraDB in terms of technical and economic scalability.



1-AryaDB market position [9]

TECHNICAL FEATURES OF AyraDB

AyraDB is a key-value database that stores data in tables on disk, where a table is a collection of records indexed by a key. The records of a table are distributed among all the servers of a cluster, based on the hash of the key that indexes the record. AyraDB implements three types of tables: fixed length, padded, and Nosql.

- Fixed-length tables feature a predefined column scheme, in which the user can configure the number of fields and their labels. All the records on the same fixed-length table have the same column structure, and the maximum length in terms of bytes of each field is explicitly defined. This type of table is particularly efficient in use cases such as time series or satellite imagery.
- Padded tables, on the other hand, have no maximum limit in terms of length of its field. This allows for greater flexibility in the amount of data stored in each record, compared to fixed-length tables.
- Nosql tables have no predefined column scheme. Each record can have an arbitrary number of fields with arbitrary labels. This makes it the most flexible out of the three table types taken into account, as it allows for the storage of unstructured or semi-structured data.

To make the database more reliable, AyraDB allows users to configure a replication factor for each table. The replication factor determines the number of copies of each record that are stored on different servers. If one server fails, the system can still operate as long as there are copies enough of each record available.
AyraDB also offers a set of HTTP-REST APIs for basic record-based operations like reading, writing, and deleting records. Moreover, it provides APIs for creating, deleting, and modifying tables, as well as for managing servers in a cluster. At the record level, AyraDB ensures consistency by synchronizing data across multiple servers. In case servers become unavailable, the write and delete operations are logged and executed when the servers become available again. AyraDB can operate consistently with up to R-1 failed servers, where R is the replication factor.

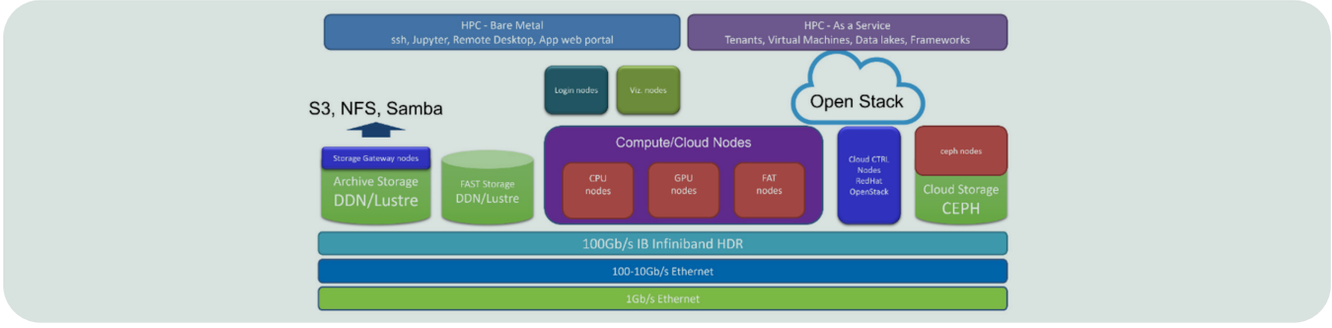
HARDWARE INFRASTRUCTURE: davinci-1

The performance tests conducted for this study have been carried out on a cloud infrastructure set up based on the OpenStack® cloud technology [8]. This cloud technology makes it easy to deploy the necessary servers required for conducting such tests.
The computing nodes used to set up the servers are part of a high-performance computing cluster named davinci-1 located in Genoa and owned by Leonardo. Such HPC cluster consists of a total of 56 CPU nodes and 80 GPU nodes. The nodes are connected to each other through both InfiniBand and 10 Gbps Ethernet connections, but only Ethernet has been used for the performance benchmarking. Figure 2 provides a detailed schematic of such infrastructure. The nodes used for the cloud tenant belong to the CPU family and are equipped with 2 24-core CPUS (Intel® Xeon® Platinum 8268 CPU @ 2.90GHz) and 768 GiB of RAM. Through the OpenStack® layer, the virtualization of the servers that are used for scaling of the benchmarks described in this work, takes place.

BENCHMARK PIPELINE

The AyraDB Benchmarking Backend

The AyraDB Benchmarking Backend (ABB) is a system created to enable easy and straightforward benchmarking of AyraDB clusters of any size and configuration. The system is designed for simplicity and ease of use, as it requires only a single command to benchmark any AyraDB configuration.
The architecture of the benchmarking system is hierarchical and consists of three levels, as it is depicted in Figure 3.
At the top level, the Benchmarking Backend orchestrates the benchmarking activity.
The **metaclient**, the machine where the Benchmarking Backend runs, receives the specifications of the requested benchmark and sends commands to the client subsystem.
The **clients** translate those commands into a workload for the servers of the AyraDB cluster. The ABB system can operate with various hardware configurations, including a multi-site distributed infrastructure.



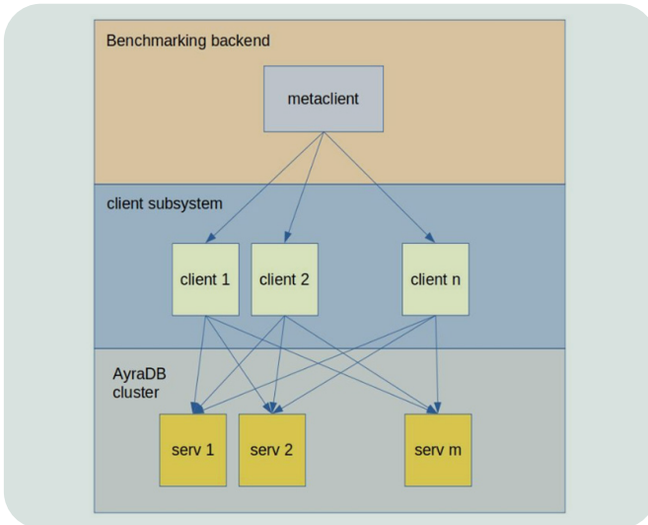
2-davinci-1 infrastructure scheme

For this benchmarking effort, the focus was on a single-site infrastructure to isolate the performance of the database from other infrastructural constraints.
Overall, the ABB system simplifies the benchmarking process for AyraDB and facilitates the interpretation of results.

Configuration of a benchmark run

The benchmark run's configuration includes the following parameters:

- M: the number of servers in the AyraDB cluster;
- B (in GBytes): the size of the table per server. For instance, with one server, the table size is B GBytes, and with M servers, it is MB GBytes;
- L (in bytes): the size of the table's records. The total number of records for the table is determined by MBx109 /L;
- C: the number of clients in the client subsystem, which follows the rule of having 1 client every three servers. For example, with 1, 2, and 3 servers, one client is used, while with 4, 5, and 6 servers, two clients are used, and so on;
- Operation type: a. Read: randomly and uniformly selects a record (or a selection of fields) from the table and sends it to the requesting client. b. Update: randomly and uniformly selects a record (or a selection of fields) from the table and modifies it according to the values provided by the requesting client;
- Addressed fields of the record (for both read and update);
- N: the number of operations (read/update) per connection per round;
- P: the pipeline size of read/update operations;
- R: replication factor;
- T: table type (fixed-length, padded, nosql).



3 - AyraDB Benchmarking Backend workflow [9]

Execution of a benchmark run

The benchmark run consists of two phases:

- Phase 1 - Initial table loading: During this phase, the clients create records and load each record onto the AyraDB cluster, to load the table on the servers.
- Phase 2 - During the measurement phase of maximum throughput for read/update, the process occurs in a series of rounds. At each round, the benchmark establishes iM connections from the clients to the AyraDB servers, where *N* read or update operations are executed on each connection. The throughput of each connection is then recorded. Once all connections have completed *N* operations, the total throughput is calculated. As the number of connections per server increases, the throughput also increases until the maximum throughput of the AyraDB cluster is achieved. When the measured throughput becomes stable, the benchmark run is concluded, and the maximum throughput is recorded. In round *i*, each server has I connections from the clients.

During the measurements of the total throughput, some bias can be introduced. Indeed, the time required for each connection to complete *N* operations can vary slightly. If the throughput of each connection is simply added up, it may lead to an optimistic bias because some connections may finish their jobs earlier than others, thus leading to larger throughput for those connections.

This bias can be avoided by evaluating the throughput when all the connections are active yielding the correct throughput evaluation. For any major details on this procedure, please refer to the paper published on this AyraDB benchmark [9]. This approach ensures that the total throughput is not biased by the completion time of each connection and provides an accurate measurement of the system's performance.

RESULTS

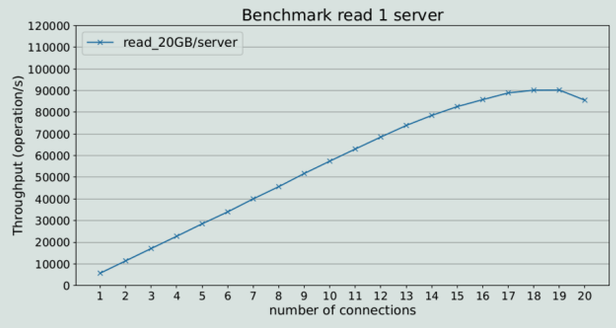
All the benchmarks utilized machines deployed using OpenStack. Both servers and clients did feature the following characteristics:

- 8 CPUs;
- 32 Gbyte RAM;
- 256 Gbyte SSD.

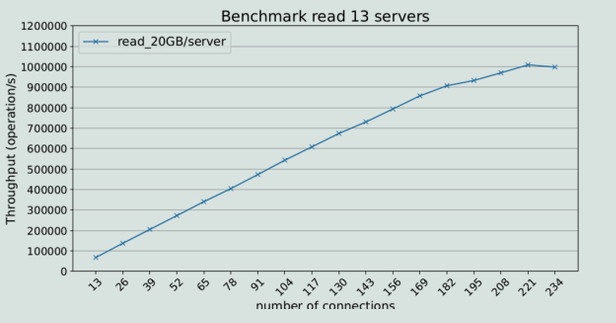
The benchmark parameters are instead:
- M, which represents the number of servers in the AyraDB cluster, ranging from 1 to 13;
- B (GByte), the size of the table per server, which we've set to 10, 20, 30, and 40 GByte/server;
- L (byte), the size of the table's records, which we've set to 1500 byte;
- C, the number of clients in the client subsystem, is a function of the number of servers.

Other parameters include:
- the type of operation: read/update;
- the addressed fields of the rec: rd one field of 100 byte;
- N, the number of operations (read/update) per connection per round: 1 million;
- P, the pipeline size: 16;
- R, the replication factor: 1;
- T, the table type: fixed-length.

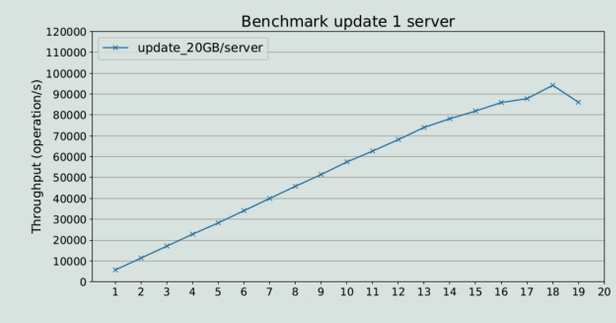
In Figure 4, the throughput of the read operation is plotted against the number of connections for an AyraDB cluster consisting of one server, and a table size of 20 GByte. The graph shows that the throughput steadily increases as the number of connections increases, until it reaches the maximum system's throughput, at which point it saturates.



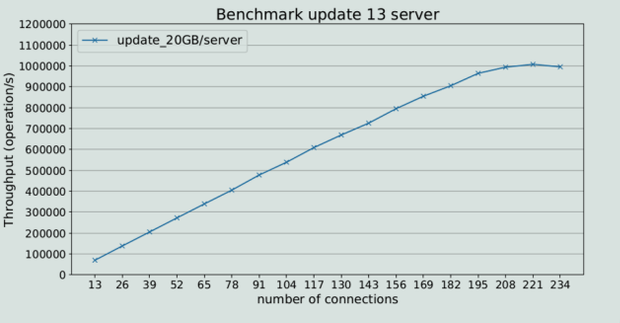
4 - Measured throughput of read operation, 1 server, table size 20 GByte/server [9]



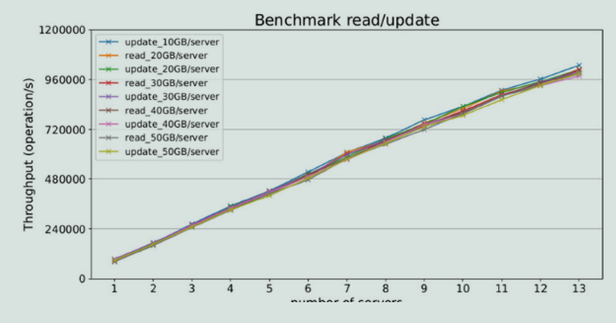
5 - Measured throughput of read operation, 13 server, table size 20 GByte/server [9]



6 - Measured throughput of update operation, 1 server, table size 20 GByte/server [9]



7 - Measured throughput of update operation, 13 server, table size 20 GByte/server [9]



8 - Maximum throughput of read/update operation, 1 to 13 servers, server, table size equal to 10 GByte/server, 20 GByte/server, 30 GByte/server, 40 GByte/server [9]

Figure 5 shows a similar trend, but on a different scale, as it reports the measured throughput of the read operation as a function of the number of connections for an AyraDB cluster consisting of 13 servers. Q ualitatively, the behaviours shown in the two figures are quite similar, but with 13 servers, the throughput exceeds 1 million read/s.
The same tests have been replicated also for the update operation (Figure 6 and Figure 7), thus obtaining results similar to those concerning the read operation, distinguishing between a cluster of 1 and 13 servers.

Finally, Figure 8 shows the scalability of the maximum throughput of both the read and the update operations as a function of the number of servers in the AyraDB cluster, for different table sizes. The results show that the throughput increases almost linearly with the number of servers, which indicates good scalability. The maximum throughput achieved for a single server is around 100,000 operations per second (ops/s) for both the read and the update operations, while for a cluster of 13 servers with a table size of 40 GByte/server, the maximum throughput exceeds 1 million ops/s for both the operations.
The results also show that the table size impacts the maximum throughput, as larger table sizes result in lower maximum throughput values. Overall, the results demonstrate the scalability of AyraDB across a range of cluster sizes and table sizes.

CONCLUSIONS

This paper presents AyraDB, a next-generation database designed to address scalability issues faced by companies with big data requirements. The slowdown of Moore's law is a root cause of current scalability issues. AyraDB offers linear scalability and per-core throughput that is 5 times greater than of the best available benchmark.
The target of the future work will be to benchmark AyraDB with Infiniband, to hopefully break the 1 million operations/s limit.

This paper also demonstrates how AyraDB's thrifty hardware consumption through greater performance translates into lower costs, and thus increases economic scalability.
The ongoing work includes the testing of AyraDB with a growing dataset size, performs a comparison with different types of tables supported by AyraDB, and builds a data processing pipeline for key applicability areas featuring stringent performance requirements.

Originally published online by CEUR Workshop Proceedings (CEUR-WS.org/vol-3340/paper37.pdf)

Roberto Morelli: roberto.morelli.ext@leonardo.com

REFERENCES

[1] IDC "Worldwide semiannual big data and analytics spending guide", 2021, https://www.idc.com/getdoc.jsp?containerId=IDC_P33195
[2] www.ayradb.com
[3] Boudriki Semlali, BE., El Amrani, C. (2021). Satellite Big Data Ingestion for Environmentally Sustainable Development. In: Ben Ahmed, M., Mellouli, S., Braganca, L., Anouar Abdelhakim, B., Bernadetta, K.A. (eds) Emerging Trends in ICT for Sustainable Development. Advances in Science, Technology & Innovation. Springer, Cham. https://doi.org/10.1007/978-3-030-53440-0_29
[4] Fuad Bajaber, Sherif Sakr, Omar Batarfi, Abdulrahman Altalhi, Ahmed Barnawi, Benchmarking big data systems: A survey, Computer Communications, Volume 149, 2020, pp. 241-251, ISSN 0140-3664

[5] Couchbase performance benchmarks, https://www.couchbase.com/benchmarks, 2022
[6] Scylladb performance benchmarks, https://www.scylladb.com/product/benchmarks/aws-i2-8xlarge-benchmark/#:~:text=ScyllaDB%20on%20AWS%20i2.,8xlarge%20Benchmark&text=Benchmarking%20database%20systems%20helps%20users,of%20wide%20column%20NoSQL%20databases.
[7] Ahmet Ercan Topcu, Aimen Mukhtar Rmis, Analysis and evaluation of the riak cluster environment in distributed databases, Computer Standards & Interfaces, Volume 72, 2020, ISSN 0920-5489, https://doi.org/10.1016/j.csi.2020.103452
[8] https://www.openstack.org/
[9] Carlo Cavazzoni, Chiara Francalanci, Paolo Giacomazzi, Nicolò Magini, Roberto Morelli and Paolo Ravanelli, Benchmarking AyraDB Next-Generation Database on davinci-1 Super-Computer, ITADATA 2022.
[10] https://www.cherry-data.com/

Published and Printed by:
Leonardo S.p.A.
Chief Technology and Innovation Office
Piazza Monte Grappa, 4
00195 Roma

The Editorial Team thanks Lucrezia Calderaro
for serving as the Guest Editor, and Paolo Casanova for his contribution.

The POLARIS Innovation Journal is an editorial initiative of the Chief Technology and Innovation Office.
Other initiatives of the POLARIS Innovation Journal are the Paperbacks and the Lunchtime Webinars.
The Journal invites questions and suggestions from readers.
Contact the Editorial Office at: polaris@leonardo.com

Scan this QR code to access the web version

https://www.leonardo.com/polaris_2023_48/

In compliance with the Leonardo sustainability policies,
and to contribute reducing the environmental footprint of the Company,
the POLARIS Innovation Journal is printed on certified paper (Xerox International Certificate).
The POLARIS Innovation Journal is published biannually.
Issue 48 – May 2023