Machine Learning
Journal Club

Annual Activity Report **2021**

Annual Activity Report 2021

Machine Learning Journal Club

Students from University and Politecnico of Turin deepen topics in AI and ML

Officially affiliated with University of Turin

Simone Azeglio, Arianna Di Bernardo, Simone Poetto, Pio Raffaele Fina, Marina Rizzi, Beatrice Villata, Gabriele Penna, Flavio Sartori, Luca Bottero, Francesco Calisto, Valerio Pagliarino, Letizia Pizzini, Enrico Sansone, Pietro Sillano (all the authors have contributed equally).

March 2022

Machine Learning Journal Club
Via Giuria 1, 10125 Torino
info@mljc.it
`https://www.mljc.it`

Foreword

MLJC was created with the aim of giving students the opportunity to get involved, to shape their ideas by putting into practice what they have learned during their studies. The stimulating and multidisciplinary environment that has been created within our organization is fertile ground for those who want to expose themselves to the multitude of stimuli that the world of research has to offer.

Many projects have been successfully concluded and others have just started. Our non-profit organization has made multidisciplinarity its pillar, thanks to the diverse backgrounds of our members. We have made machine learning tools available to the social sciences, neuroscience, physics, climate in a practical, hands-on manner. Collaborations with leading universities and top-notch companies, both national and international, have been possible thanks to the willingness of individual members who have dedicated hours of work within the organization.

We would like to personally thank each and every one of the participants in the MLJC, who have made possible the realisation of such an ambitious and challenging project.

Special thanks to the professors and collaborators of the University of Turin who supported us from day one, among them: Professor Michele Caselle for providing us with the right advices to pursue our activities, Professor Piero Fariselli and Matteo Osella for technical suggestions and active involvement. Thanks to Monica Rinero for providing us with logistic support and space to carry out our activities within the Department of Physics at the University of Turin. Thanks to Marco Aldinucci, Marco Beccuti and Claudio Mattutino for allowing us to actively use essential computational resources within the HPC4AI center, and Professor Cristina Gena for giving the Medical AI team some useful advice from their own experience. We would also like to show our gratitude to Christopher Rackauckas for giving us the opportunity to take an active part in the SciML framework, within the Julia programming Language. We are very grateful to Don Luca Peyron for his great support to our association and for providing us with everything we needed, and to NPO Torino srl, in particular to Romualdo Delmirani, Francesco Dipietromaria and Massimo Altamore for the acquisition of hardware components, relevant to several projects.

Simone Azeglio
Research Engineer at Institut Pasteur
MSc graduate at University of Turin

Arianna Di Bernardo
Research Engineer at École Normale Supérieure
MSc graduate at University of Turin

Contents

1 About the Association

We are a group of students from the University of Turin and the Polytechnic University of Turin, mostly from a STEM background. Our passion for machine learning and artificial intelligence lead to the creation of an open and non-profit organization with the aim of closing the gap between being a student and starting a research career. The MLJC is an interactive community, a coding playground where to share our geeky ideas and our expertise. We are actively engaged in research projects with a broad, interdisciplinary focus and participating in hackathons and competitions related to AI. The aim of our association is to promote hands-on knowledge of machine learning and artificial intelligence across the academic and social environment of Turin.
We believe that interdisciplinarity and cross-contamination can be the key to new discoveries and fruitful, engaging research projects. We try to innovate learning and research paradigms with a focus on self-organization, peer-learning and peer-reviewing, bridging the gap between university courses and real-world applications and ultimately producing reusable open source software that have an impact on society, industry and science.
In order to promote machine learning research we jump-started a wide set of projects, on various premises and online:

- We took over the University of Turin after-hours, delivering free-for-all courses on Python for scientific programming and machine learning
- We engaged our more experienced members in open and horizontal research projects on natural language processing, theoretical machine learning, scientific machine learning and machine learning for biomedical applications.
- We encouraged and supported the participation of our members to hackathons and competitions.
- We published divulgatory articles on machine learning techniques and hot topics, some of our best works lead to publications, conference talks and pre-prints.

1.1 Our story

We started our first activities as a little group of students from the M.Sc. in Physics of Complex Systems at the University of Turin. We were curious about emergent possibilities stemming from Data Mining, Machine Learning and Artificial Intelligence: that's why we started experimenting with real-world projects and participating in some Kaggle's competitions we were passionate about (for example, we took part in the Deep Fake Detection Challenge).
Going forward, we decided to open our group to students from different backgrounds: in order to build some common knowledge, we started organizing extra-curricular lessons and activities, covering Python applied to machine learning.
As we meet and engage with new people from all walks of science, our community grows large and diverse: we are proud to count members from other branches of physics, computer science, engineering, mathematics, biology, economics and neuropsychology to name a few. We meet regularly (online or in person), to work and organize projects related to machine learning and its applications.

2 Meet the team

Simone Azeglio - President

Simone is currently a research engineer at the Hearing Institute, Institut Pasteur, in Paris, working on machine and deep learning activity-driven models for auditory perception. He is a Physicist by training, previously enrolled in the M.Sc. in Physics of Complex Systems at the University of Turin (Italy), with a keen interest in Representation Learning. More in detail, he is interested in unveiling the role of symmetries and equivariance in neural (biological and artificial) representations, fundamental for designing better inductive biases and data efficient architectures.

Arianna Di Bernardo - Vice president

Arianna is a research assistant in Theoretical Neuroscience at École Normale Supérieure, in Paris. She graduated in Physics of Complex Systems at the University of Turin.
She is interested in modeling complex brain functions and in discovering low-dimensional structures in neural data, through geometric and topological tools. Moved by this interest, she contributed to establish the research line of *Medical AI and Brain Computer Interfaces* within the Machine Learning Journal Club.

Francesco Calisto

Francesco is in his third year of the Bachelor Degree in Physics at the University of Turin. He is mainly focused on the theoretical and mathematical aspects of physics. He also conducts researches in machine learning, with a particular interest towards the connection between machine learning and physics. Recently he has worked in the field of Scientific Machine Learning with state-of-the-art Julia libraries, such as NeuralPDE.jl.

Marina Rizzi

Marina is a PhD student in Economics at Collegio Carlo Alberto and University of Turin.
She is an Applied Economist, with a focus on Political Economics, Public Economics and Media Economics. She is interested in machine learning methods applied to economics and social sciences. In particular, she tries to use methods of text analysis and natural language processing techniques to shed light on topics related to her research.

Micol Olocco

Micol is enrolled as intern at the University of Geneva (Switzerland). She is currently working on model based approaches for anomaly detection in large-radius jet. She graduated in Nuclear Physics at the University of Torino (Italy). Her thesis project involved Machine Learning applied to Natural Language Processing for analyzing the error messages generated during the file transfers across the CERN grid.
Thanks to her thesis project, she got in touch with Machine Learning field and she is fascinated by it. She believes data are our instrument to understand the present, learn from the past and get close to the future. She is curious, greedy of learning and likes new challenges.

Luca Bottero

Luca is an undergraduate student in Physics at the University of Turin. He is interested in many aspects of Machine Learning, in particular those of relevance for physics or science in general.
In particular, Luca thinks that NLP has the potential to help scientists doing research better and faster. In addition, he explores the theoretical aspects of ML to better understand its strength and vulnerabilities.

Pio Raffaele Fina

Pio is a Msc student of Artificial Intelligence and Information Systems at University of Turin.
He is keen on digital signals, especially image processing. His main interest lies in understanding how those signals can encode information patterns and semantic structures of the complex reality around us.
Currently he is working as junior machine learning engineer in biomedical applications.

Valerio Pagliarino

Valerio is a student in Physics (Bachelor) at the University of Turin. He is interested in electronics, embedded systems, computing and signal processing as tools for investigating physics, such as the technology of the detectors employed in high energy physics. Considering that Machine Learning techniques are gaining importance in these fields, he is interested in deepening his knowledge on ML architectures that can support physical simulations, data analysis and related applications.

Letizia Pizzini

Letizia is a physicist, currently employed as a PhD student in Complex Systems for Quantitative Biomedicine at the University of Turin. She mainly focuses on Systems Biology study through the application of machine learning techniques born to be used in different context, such as Natural Language Processing and Network Theory.
Gaining information from complex biological or medical data sets by selecting a small number of latent variables is her primary challenge.

Gabriele Penna

Gabriele is a Master's student in Biomedical Engineering at the Politecnico di Torino and a Research Assistant in the NBE department at Aalto University. He is interested in Biomechanics, Neuroengineering, and Machine Learning applied to the biomedical sciences. In particular, he is curious about the connection between Neuroscience and the Engineering world, with special attention on how the upcoming signal processing and machine learning techniques can help revolutionize this field.

Alessio Borriero

Alessio Borriero is attending the Master's Degree in Physics of Complex Systems at University of Turin. Since he was a child he has always been fascinating by stars, atoms and everything could stimulate his thirst for discovery. He is also a very competitive person and he put all of himself in everything he does, from sport to study.

Flavio Sartori

Flavio is a junior data scientist and a student of the MSc in Physics of Complex Systems at University of Turin.
He is interested in Medical ML and in Machine Learning applied to graphs, and in particular in the explainability of GNNs.

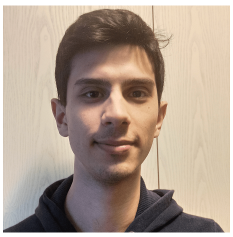
Aurora Micheli

Aurora is a Master's student in Physics of Complex Systems at the University of Turin. She is mainly interested in the fields of Computational Neuroscience and Brain Computer Interfaces, with a focus on the machine learning implications and techniques applicable to these topics.

Simone Poetto

Simone is a Ph.D. student in Computational Neuroscience at the Nicolaus Copernicus University in Torun (Poland) and ISI Foundation in Turin (Italy). After a Master degree in Physics of Complex Systems, he focused his research interest on the interplay between Neuroscience and Artificial Intelligence. In particular, he is interested in discovering how the brain creates and manipulates abstract representations of the external world. He tries to tackle these questions using tools from geometry, topology, and complex networks. He also has experience in more practical applications of Machine Learning in Neuroscience, ranging from data analysis of neuroimaging datasets to the design of better bio-inspired Artificial Neural Networks.

Andrea Semeraro



Andrea recently obtained a Master degree in Physics of Meteoclimatic System. Environment and Space are the topic that most develop his interest for the nature and his curiosity thanks also to the trainership made in Altec.
He is also intrigued by Neural Network and Multi Agent Systems Approach. Currently, he is dealing with machine learning applied for climate problems using Physics Informed Neural Network.

Pietro Sillano

Pietro is a student of the MSc in Physics of Complex Systems at University of Turin.
His main interests of research are Physics of Living Systems and Non-Equilibrium Biologic Systems using theoretical and computational methods from Physics. In addition he is interested in Brain Computer Interfaces and how Machine Learning techniques can be applied in this field.

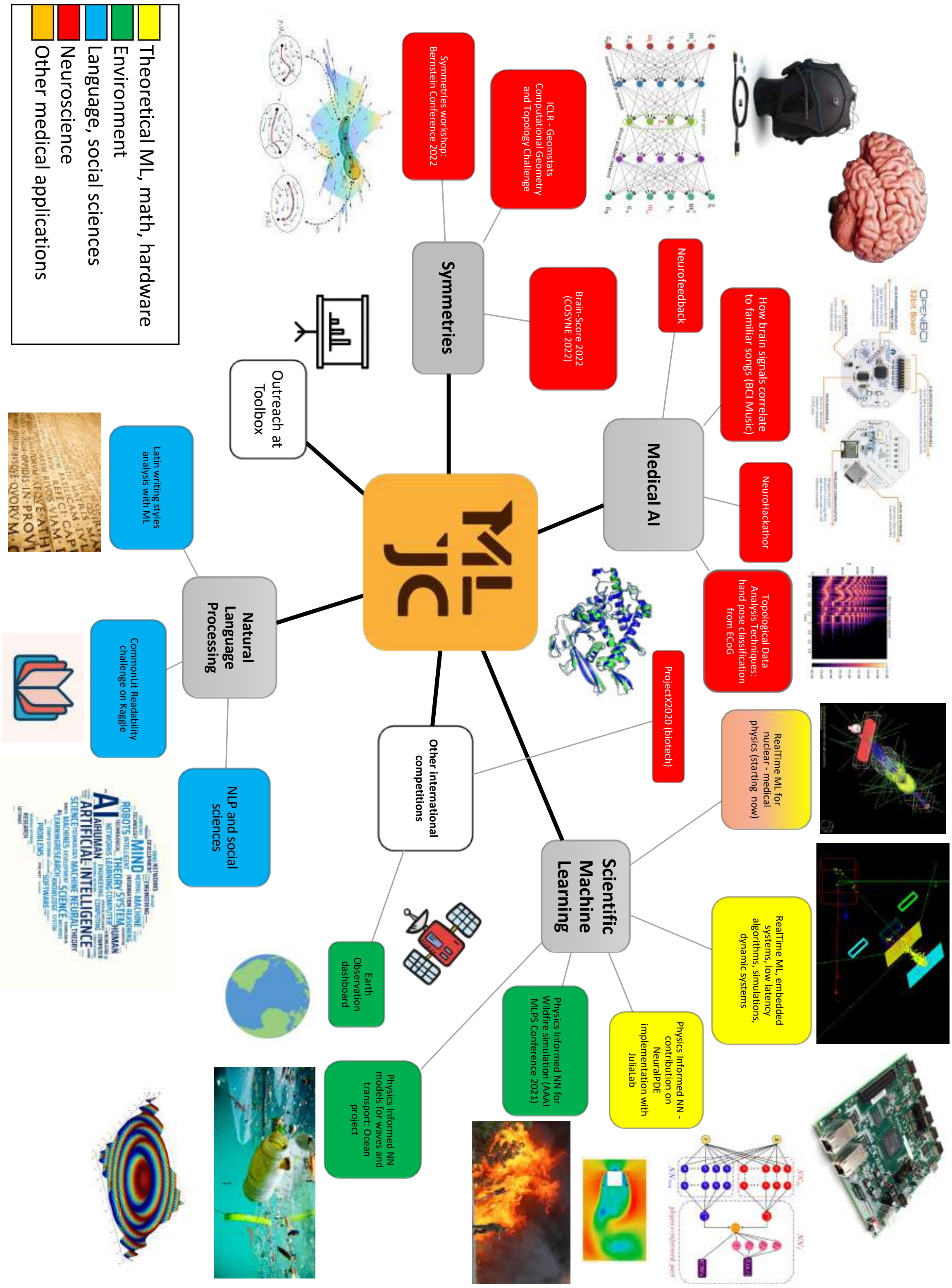
Beatrice Villata

Beatrice is at the first year of her master's in medical physics at the Ludwig Maximilian University of Munich. Here, she is involved in the different applications that physics plays in medicine. She is also interested in Neuroscience, with a focus on machine learning. With the MLJC, she took part in various hackathons about Brain Computer Interfaces.

Enrico Sansone

Enrico is a Master student in Physics of the Complex Systems at the university of Turin.
Having followed his interests in the Neuroscience field, his main topics of research are the possible theoretical representations of the processes being Decision Making, and how Deep Learning could help and be helped by this particular field of study.

3 Selection of Scientific Projects carried out in 2021



Selected Projects

Biomedical applications
with particular focus
on Brain Computer Interfaces

3.1 Medical AI and BCI

The *Medical AI and BCI* group at the MLJC aims to investigate the possibility to develop advanced machine learning tools to analyze electrophysiological data from the brain. In particular, we are carrying out several projects in the Brain Computer Interfaces (BCI) area, dealing with electroencephalography (EEG) and electrocorticography (ECoG) signals from the brain.

We are collaborating with **g.tec medical engineering GmbH** (www.gtec.at), leading company in the BCI world with which we sponsor and organize Neurotechnology and BCI training programs and conferences.

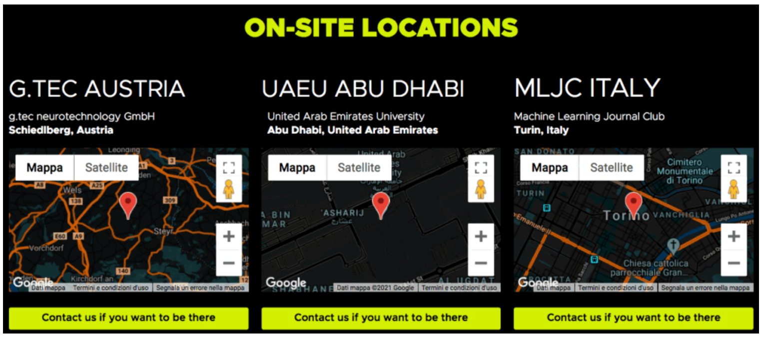
- **Spokespersons**: Arianna Di Bernardo, Gabriele Penna
- **Participants**: Arianna Di Bernardo, Gabriele Penna, Letizia Pizzini, Pio Raffaele Pina, Simone Poetto, Beatrice Villata, Flavio Sartori, Enrico Sansone, Pietro Sillano, Simone Azeglio.
- **External collaborators**: Sara Giganti, Ilaria Gesmundo, Vincenzo Triglione, Marco Bottino, Ivan Decostanzi, Nibras Abo Alzahab, Giulia Sanguedolce, Elios Ghinato, Pietro De Luca, Davide Mattioli, Federico Nardi, Alessia Camasio, Filippo Di Fortunato, Francesca Dalla Mutta, Davide D'Adamo, Michele Romani.

3.1.1 Participation to the *virtual BR41N.IO hackathon*

We participated in three editions of the hackathon organized by **g.tec** regarding EEG and Ecog data analysis:

- `https://www.br41n.io/Toronto-2020`
- `https://www.br41n.io/Spring-School-2021`
- `https://www.br41n.io/IEEE-SMC-2021`

For the last two editions the Machine Learning Journal Club has been the Co-Host of the event.



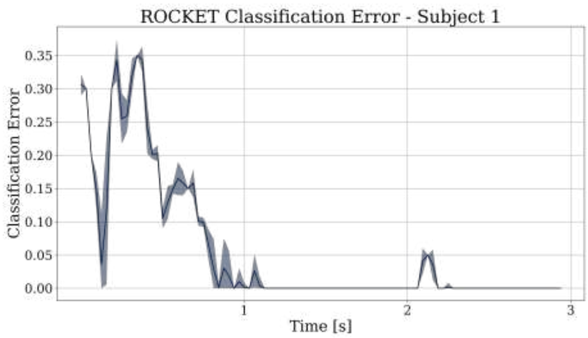
BR41N.IO is a brainstorming and collaborative marathon designed to be a learning experience for developers, technologists, engineers, students, artists, and scientists who cram and build brain-computer interface (BCI) applications together in teams. During these hackathons we had the opportunity to participate in different projects in two areas: programming and data analysis.

1. The first area involved the use of an EEG helmet to carry out an experiment chosen from those presented;
2. The second area instead envisaged receiving pre-processed data to allow competitors to develop algorithms capable of solving a specific classification task.

Some examples of the different challenges and solutions proposed by our groups during the competition are as follows:

- **Stroke Rehab Data Analysis**: We analyzed a motor imagery BCI data-set from a chronic stroke patient. BCI systems can be combined with different types of external devices to assist the execution and learning of movements. In the approach for movement restoration, stroke survivors perform MI exercises while wearing an EEG cap. The decoded brain oscillations can be used to move a VR avatar and/or trigger another feedback mechanism to reproduce the imagined movement.
  Our initial idea was to separate the different classes through the Common Spatial Pattern (CSP) technique, maximizing the difference of the variances of the signals between the left and right trials. In the end, instead of using a linear classifier we decided to use an unsupervised technique of manifold learning (UMAP). For the classification step we decide to make use of a Multi-Layer Perceptron.
- **SSVEP Data Analysis**: We analyzed a SSVEP BCI data-set from a healthy person. An EEG evoked potential (EP) is a distinctive pattern of positive and negative voltage deflections that is time-locked to a specific sensory stimulus or event. Visually evoked potentials (VEPs) are those evoked by sudden visual stimuli, such as a light flash, the appearance of an image, or an abrupt change in color or pattern. Steady-state VEPs (SSVEPs) are stable oscillations in voltage that are elicited by rapid repetitive stimulation such as a strobe light, a light-emitting diode (LED), or a pattern-reversing checker box presented on a monitor.

We developed our work in three steps: preprocessing, Feature Extraction and Classification. During the preprocessing we applied a temporal filter to eliminate slow drifts, a Common Averege Reference (CAR) spatial filter and a Independent Component Analysis (ICA). For the Feature Extraction the ROCKET algorithm (RandOm Convolutional KErnel Transform) has been used. In the end we trained a ridge classification model for each class applying one versus rest classification, with a L2 regularization. We evaluated the performance of the classifier using 5-fold Stratified Cross Validation. We used 20 trials for the training set (training with cv) and 20 for the test set.



- **ECoG Hand Pose Data Analysis**: During the Spring School 2021 competition one of the group worked on an ECoG BCI data-set from an epilepsy person winning the hackathon.

The work done during the hackathon has been further developed leading to the drafting of a paper entitled **"Topological Data Analysis (TDA) Techniques Enhance Hand Pose Classification from ECoG Neural Recordings"**.

3.1.2 Invited Talk at Intesa Sanpaolo

The Machine Learning Journal Club has been a guest at the **Neuroscience Lab** of **Intesa Sanpaolo Innovation Center** on Wednesday 27 October, from 11 am to 1 pm, presenting a seminar entitled **"Brain Computer Interface: a new future for disabilities"**.

The seminar, held by Gabriele Penna, Pio Raffaele Fina and Pietro Sillano (members of the MLJC) is part of a series of events organized by the Neuroscience Lab of Intesa Sanpaolo Innovation Center which aims to illustrate and disseminate the evolution of research conducted in the field of Neuroscience.

Abstract: A Brain-Computer Interface (BCI), is a connection between the brain and an external device capable of translating and directing brain signals to an external output, such as the control of a cursor or a prosthesis.
Thousands of young people with physical and cognitive disabilities live, grow and learn with limited communication and interaction with the world around them. BCIs are a promising means of helping them with the challenges they face, equipping them with effective tools to improve their understanding and communication skills. The scientific challenges are quite a few, in particular with regard to the management and analysis of the data collected, and innovative technical data analysis and machine learning activities are increasing.
Through their projects and competitions, the students of the Machine Learning Journal Club (MLJC) accompany us in the world of BCIs, ranging from how the latter can revolutionize the museum experience of people with cognitive disorders up to restoring movement

and communication skills to quadriplegic patients.

The seminar also led to the recording of an episode of the podcast **"Innovation Coffee e seminari dell'innovazione"**, the column edited by Intesa Sanpaolo dedicated to the dissemination of the culture of innovation, which deals with issues inherent to the trends of the new millennium. The episode of the podcast can be found at the following links:

- **Spotify Podcast**:`https://tinyurl.com/37tzv2w2`
- **Google Podcast**: `https://tinyurl.com/2trw8mx9`
- **Apple Podcast**: `https://tinyurl.com/2vrnumjw`
- **Intesa Group Website**: `https://tinyurl.com/nhhkecr9`

3.1.3 Participation to the *NeuroHackathor 2021* hackathon

In the spring of 2021, some members of MLJC participated in the 1st edition of *Neurohackathor*, a 24 hour hackathon hosted by the *Neurotechnology Scientific Student Club* from Torùn, Poland. NeuroHackathor 2021 is an international hackathon for engineers, developers and cognitive scientists interested in neurotechnology and human daily living. We brought 11 people organized in 2 teams working on two different topics.
The first team worked on the "*Museum of the future*" topic, developing the idea for an hybrid BCI-NLP system (Figure 1). The aim was the **customization** of the museum visitors' experiences through cutting edge technology; using multiple bio-signals as EEG, eye tracking, galvanic skin response and NLP technologies like *automatic summarization*. The proposed system is able to adaptively modify the caption of a picture in response to the engagement of the viewer. A further extension of this work has been presented to a wider audience during a seminar held at the Intesa Sanpaolo's Innovation Center.

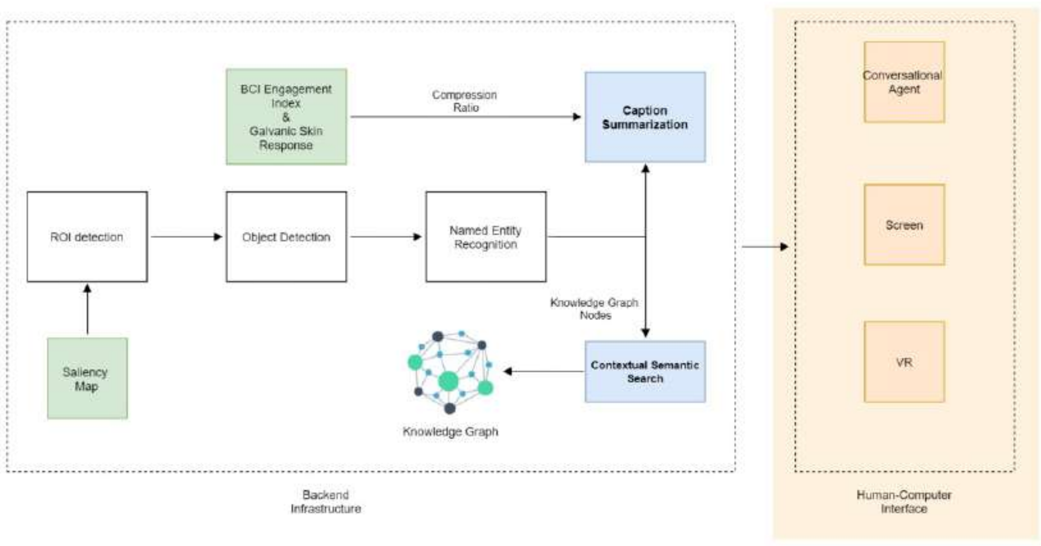


Figure 1: Component diagram of the proposed BCI-NLP system

The second team chose to work on the topic: "*How to reduce the negative effects of pandemic isolation with neurotechnology?*". The main idea was to simulate the feeling of being in the nature, which is proven to be beneficial for depression, in a situation where a subject is forced in a closed space. They proposed a **neurofeedback protocol** and designed a minimalist & easy-to-use device, "The EGG", to perform the protocol (Figure 2).
Both teams were awarded, respectively, with 2nd and 3rd place. Besides this achievement, we started a fruitful collaboration with the Neurotechnology Scientific Student Club.

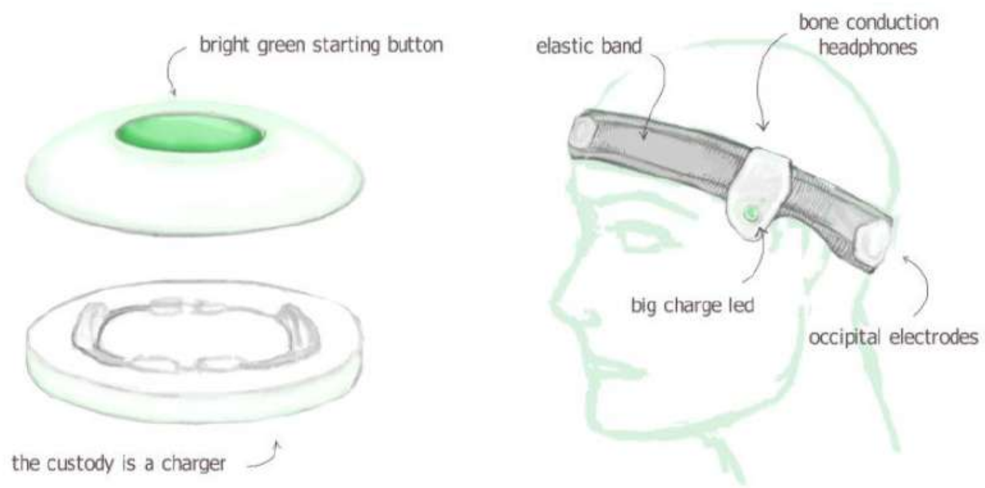


Figure 2: Concept design sketch of "the EGG", the proposed device.

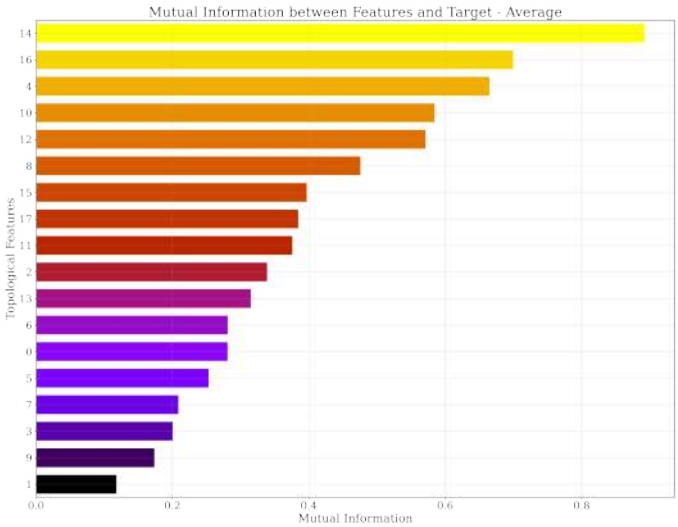
3.1.4 Drafting of the paper *Topological Data Analysis (TDA) Techniques Enhance Hand Pose Classification from ECoG Neural Recordings*

Pre-print: `https://arxiv.org/abs/2110.04653`
GitHub Code: `https://github.com/MachineLearningJournalClub/ECoG_VBH_2021`
Status: **Completed**

Electrocorticogram (ECoG) well characterizes hand movement intentions and gestures. In the present work we aim to investigate the possibility to enhance hand pose classification, in a Rock-Paper-Scissor - and Rest - task, by introducing topological descriptors of time series data. We hypothesized that an innovative approach based on topological data analysis can extract hidden information that are not detectable with standard Brain Computer Interface (BCI)techniques. To investigate this hypothesis, we integrate topological features together with power band features and feed them to several standard classifiers, e.g. Random Forest, Gradient Boosting. Model selection is thus completed after a meticulous phase of bayesian hyperparameter optimization. With our method, we observed robust results in terms of ac-curacy for a four-labels classification problem, with limited available data. Through feature importance investigation, we conclude that topological descriptors are able to extract useful discriminative information and provide novel insights.Since our data are restricted to single-patient recordings, generalization might be limited. Nevertheless, our method can be extended and applied to a wide range of neurophysiological recordings and it might be an intriguing point of departure for future studies.

The current work was presented at the Virtual Brain Hackathon organized by g.tec in April 2021, and subsequently at the *BCI & NEUROTECH MASTERCLASS ITALIA 1.0* (https://www.gtec.at/it/masterclass-italy-2022/), in February 2022.

We acknowledge the University of Turin's and Polytechnic University of Turin's High Performance Centre for Artificial Intelligence (HPC4AI) for providing us with the following computational resources: 12 Intel Xeon vCPUs, 20 GB RAM, NVidia T4 GPU.



3.1.5 Neurofeedback training to improve concentration*

Neurofeedback training, also called *neurotherapy* or *neurobiofeedback*, is a non-invasive technique, based on principles of operant conditioning. This method involves electrodes placed along the individual's scalp that transduce brain activity at predetermined locations. When a specific form of brain wave activity is recorded, a reward is offered to the subject. Typically, rewards are offered when the amplitude of a specific frequency band is measured above or below a predetermined threshold. The most commonly employed rewards are video or audio signals, either singly or in combination.

Over time, an individual can be trained to regulate his or her brain oscillations and maintain the amplitude of these oscillations within a predetermined bandwidth. It has been applied in cognition, sports, depression, attention deficit hyperactivity disorder (ADHD), Alzheimer's disease and memory.
Previous studies have reported that participants can modulate and increase the EEG band power via NFB training, resulting in improved concentration, attention to detail, and ability to enter into the flow state. In particular, concentration could be markedly improved by NFB training but many of the papers written about this application focus on the oscillations of only one (or two) band, setting thresholds for the return of negative reinforcements or rewards and often they lack significant results or do not take into account a proper sham

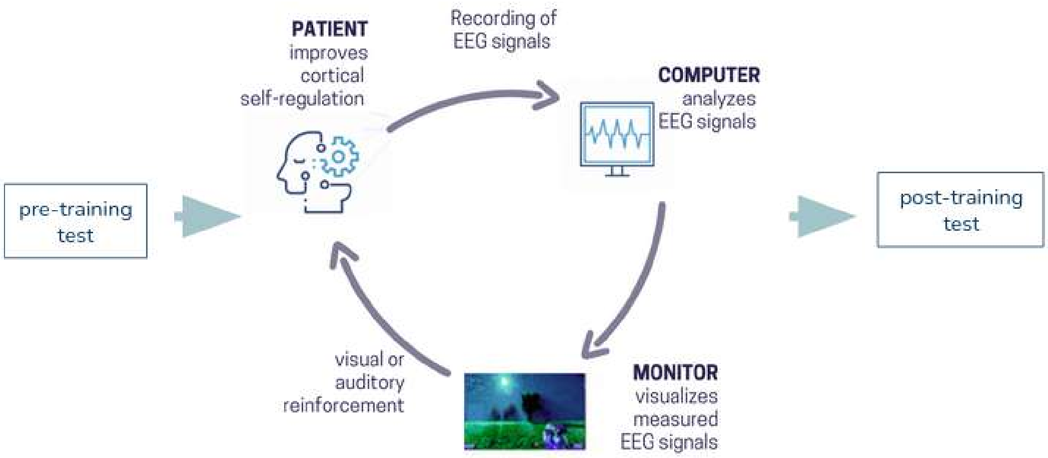


Figure 3: Neurofeedback loop design.

group to highlight eventual placebo effects in the treatment.

Moving away from the previously cited studies, our project aims to go beyond the observation of oscillations of a single band, defining a pattern (and not a single variable) on which relying to return the feedback. We start with an assessment to gain a better understanding of how brain performance is contributing to a specific mental state (as concentration or no-concentration). Then, the main point consists of the implementation of Machine Learning techniques on online datasets to extract a pattern of state-related features, even with a spatial organization. So we aim to obtain a set of features that allow modeling a neurofeedback software that returns rewards based on the entire brain configuration.

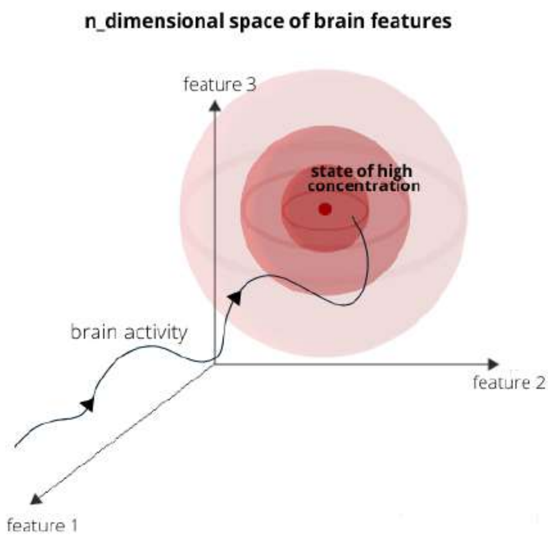


Figure 4: n_dimensional space of brain features. Concentration mental state can be represented as a point in the n_dim space, around which thresholds are chosen (displayed as concentric spheres) corresponding to increasing level of concentration that define the feedbacks returned during the training.

Therefore the training part of the project, consequential to the Machine Learning pipelines on online data, is divided into four stages: the concentration tests will be submitted to the study participants during the first and third stage; the second stage consists of the neurofeedback training, where the candidates will receive various training sessions following a specific protocol.
Protocol drafting and the evaluation of the test results is supervised with the help of psychologists.
The last step of our project includes the analysis and classification of the recorded signals to highlight, with the support of the tests performances analysis, if and how our NFB based on a complex pattern of brain features can constitute a better training for concentration than those previously tested.

3.1.6 Do you know it? How brain signals correlate to familiar songs*

Analyses of EEG signals during music perception have been conducted by many authors and in many different forms. The main difficulty is to design an experiment that allows to investigate a specific cognitive process while disentangling it from the others that are going on. In fact, listening to music elicits a broad range of cognitive processes (auditory sensations, emotions, memory retrieval, pattern recognition and prediction). In turn, those processes are modulated by many different factors: personal, cultural, and neurobiological. Also at the level of signal recording, there are different approaches. On one side it is possible to record single channels or brain areas, to measure changes in spectral power or Event-Related Potential (ERP). On the other side, one can record from the whole brain with multiple channels and study the functional connectivity that emerges after defining some sort of correlation between channels.
In this study we want to focus on familiarity with music, analyzing the differences in brain activity while listening to familiar songs compared to new ones.
The first phase of the project is designing an experimental protocol that allows to focus on the effect of familiarity while ruling out all other effects that can influence music perception. To do so we are creating a unique data set composed of dozens of musical excerpts (duration around 10 seconds) from different music genres. Music pieces are separated by some seconds of silence, during which a hypothetical subject can self-report his familiarity with the song, using a discrete scale.

Recording the brain activity while listening to those songs for a sufficient long time (10 seconds or more) should allow us to conduct different analyses on the signal. In this way we hope to study both the first reaction to familiarity (which should occur in the very first seconds of listening, as soon as the subject feels the common reaction that happens when perceiving something familiar), as well as the sustained reaction to this, that should involve other areas of the brain responsible for memory retrieval.

3.1.7 Comparison between CQT and CIT methods for EEG signal processing for information recognition*

How the human brain is able to store and recollect memories is one of the most studied topics in Neuroscience and Psychology field, whether it be how they are formed in young age, or how they can be restored in old people or Alzheimer patients. Among all the information the brain can register, it can also classify them as ”familiar”, or as something that holds an important role in our lives, such as our address, our car keys, or the face of loved ones.

To study whether something has significance for a person, two are the methods mainly used: the Comparison Question Technique (CQT) and the P300-based Concealed Information Test (CIT). The first method is based on a series of simple binary answer questions, observing and analyzing patterns in the EEG signals while the patient is been inquired. This method has been declared ”weak” and subsequently considered inferior to the second method, the CIT, which relies on finding and analysing the event related potential (know as P300) which manifests inside the brain when the patient recognizes something they know about or is somehow familiar to them. However, the comparison between the two has always been made between very different experiments, with different protocols and samples and sometimes even after many many years.
This study proposes itself as a bridge between these methods, exploring exactly why one is preferred over the other, find if it's possible to combine them together, and, as our ultimate goal, to have a clear theoretical understanding on what are the signals involved when the brain recognizes important information. The focus will be on the development of the algorithms used, and only later the drafting of an experimental protocol, in order to implement our findings. The project is divided into 3 stages:

- Development of the CIT algorithm using the recent state of the art procedure, aiming to master it as much as it concerns our study.
- Development of the CQT algorithm in relation to the work previously done with the CIT method.
- Drafting of the experimental protocol under the supervision of psychologists, in hope to be able to gather EEG data, and the final implementation and test of our work.

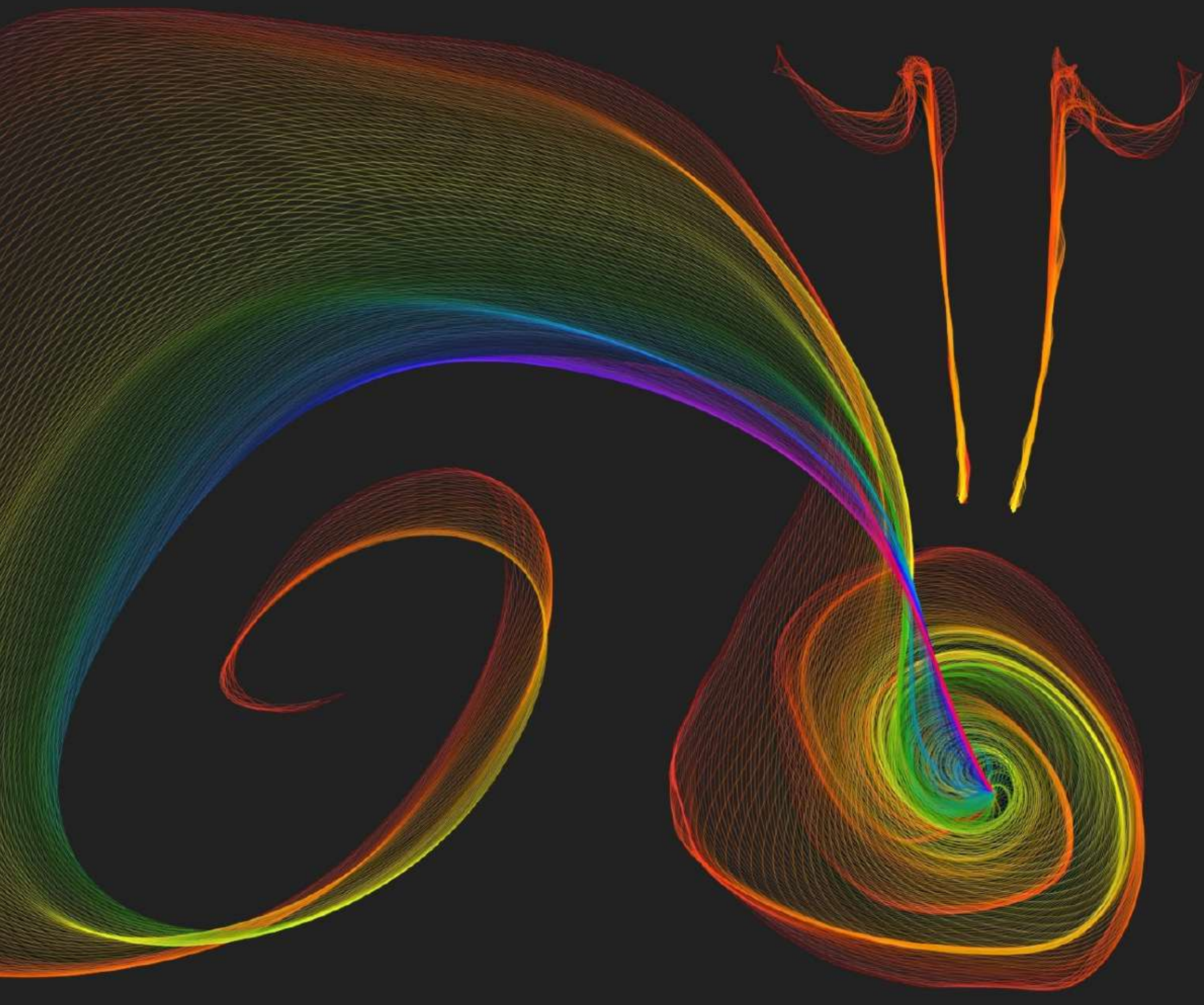
This type of procedure could help cast some light on how memories are recollected and classified. To list an example of what we have in mind, it could let us understand if an old person affected by Alzheimer retains some type of unconscious memories, related to something they perceive as familiar, and work from there to try restoring them.

---

*Projects marked with * are proposals for the next academic year. We are working on the implementation of the research lines mentioned above, and thanks to the generous contribution of NPO Torino, we have the necessary equipment to carry out the research. Subject to acceptance by the University of Turin's ethics committee, the projects will be developed in the following months.*

Selected Projects

Scientific Machine Learning



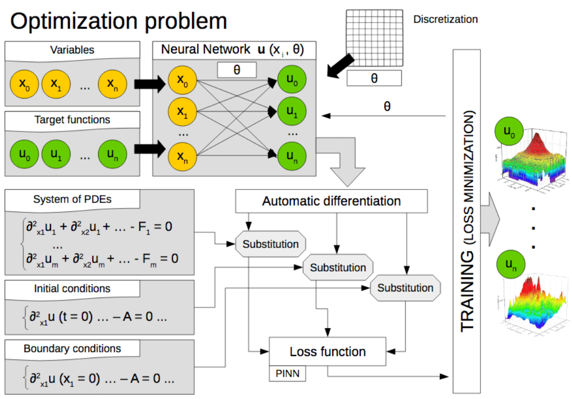
3.2 Scientific Machine Learning

The aim of this group is to explore techniques that combine machine learning models with numerical and physics-based simulations to tackle a wide range of problems in scientific disciplines. Some projects include embedding information from PDE systems inside ML models, modeling of dynamical systems, application to control systems, deployment on embedded electronics.

- **Spokespersons**: Francesco Calisto, Valerio Pagliarino
- **Participants**: Francesco Calisto, Simone Azeglio, Valerio Pagliarino, Luca Bottero, Beatrice Villata, Vittoria Berta
- **External collaborators**: Martina Scauda, Sara Tiengo, Giovanni Graziano, Mattia Falco, Francesca Bellucci.

3.2.1 Participation to the ProjectX-2020 competition on ML applied to climate change: Physics Informed Machine Learning Simulator for Wildfire Propagation

AAAI Proceedings: `http://ceur-ws.org/Vol-2964/`
GitHub Code: `https://tinyurl.com/mljc234523`
Status: **Completed**
Participants: Francesco Calisto, Valerio Pagliarino, Simone Azeglio, Luca Bottero, Martina Scauda, Sara Tiengo, Giovanni Graziano



Between 2020 and 2021 some of the Scientific Machine Learning group members, in collaboration with Martina Scauda, Sara Tiengo and Giovanni Graziano, mathematics and physics students at the University of Turin, started the project "Physics Informed Machine Learning Simulator for Wildfire Propagation" as a proposal for the *ProjectX-2020*, an international competition organized by the *UofT AI*, an association of the University of Toronto. The focus was on machine learning-based strategies for tackling climate change problems. In this competition our group got in touch with top experts in the field of climate modeling and environmental physics. The level of the competition was very high, with the participation of top American universities, such as the MIT, Carnegie Mellon University, California Institute of Technology and Stanford and our work got a very positive evaluation by the international jury.

As stated in the paper abstract, the aim of this work is to evaluate the feasibility of re-implementing some key parts of the widely used Weather Research and Forecasting WRF-SFIRE simulator by replacing its core differential equations numerical solvers with state-of-the-art physics-informed machine learning techniques to solve ODEs and PDEs in order to increase computational efficiency on high-dimensional problems and allow for future introduction of data-driven components. The goal is to evaluate the feasibility of a real-time simulator for wildfire spread prediction based on PINNs. Our ML approach is based on Physics Informed Neural Networks implemented in the NeuralPDE.jl library, which turns an integration problem into an optimization one by approximating the PDE solution with a neural network. Furthermore, this approach can successfully tackle ill-posed problems and it is more efficient for high-dimensional PDEs with respect to conventional numerical solvers.

In particular, for this project, we applied this architecture to tackle the resolution of the level-set equation, that is used to model the fire line propagation. The partial differential equation has the following form:

$$\partial_t \psi + S||\nabla\psi|| = 0, \tag{1}$$

More details regarding the hyperparameters of our model as well as the results we obtained on a simple wildfire simulation are shown in the following table and plot (Fig.5)

Optimization Algorithm	ADAM
Iterations	4800
Final Objective Value	6.39 e-8
Training Strategy	QuadratureTraining()
Domains	$t \in [0,10]$, $x \in [0,10]$, $y \in [0,10]$
Training Mesh Size	$[dt, dx, dy] = [0.17, 0.02, 0.02]$
Boundary Condition	$u(0,x,y,\theta) = ((5(x-0.3))^2 + (0.15y)^2)^{\frac{1}{2}} - 0.2$
Neural Network dimensions	$3 > 16 > 1$
Training Time	647 s

The PINN model was able to capture complex dynamic such as the asymmetric flame propagation due to the direction of the wind, which was given among the input data. On the other hand, it is possible to appreciate a discrepancy between our prediction and the form of the fireline forecasted by WRF-SFIRE because we had to assume the fuel map of the territory to be constant for technical reasons. In particular, at that time the NeuralPDE.jl library had some limitations when importing matrices of measured data for constraining the solution.

The neural network training time, that replaces the integration time of a traditional numerical problem, although it cannot be directly compared with the WRF-SFIRE benchmarks, shows the potential of this approach for tackling non-linear high-dimensional problems that occur in climate physics. Further improvements include a better integration of the physical model with the data-driven metrics during the training and a more complete problem modeling that incorporate coupling with the atmospheric simulation engine. In our paper a real-world test of our model, based on the Isom Creek fire in 2020 (Alaska) is showed, as well as a more detailed description of the model theory and implementation.

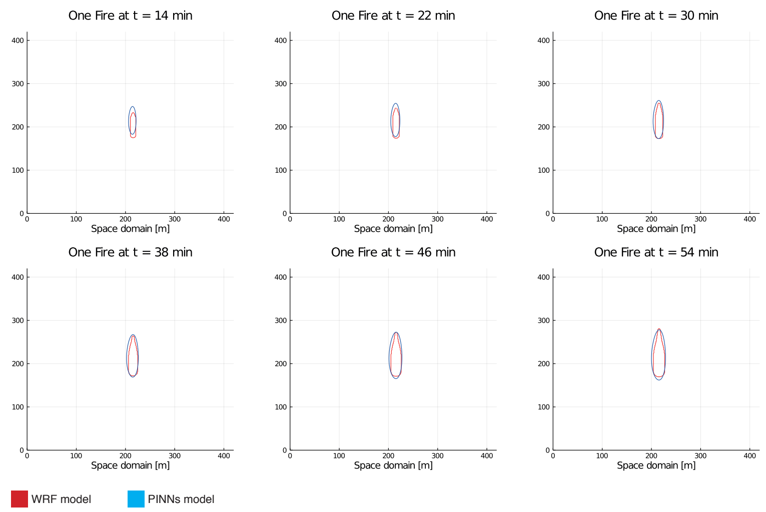


Figure 5: Comparison between the WRF and PINNs outputs at different time frames

Our team had the honour to present this work at the 2021 AAAI Spring Symposium on Combining Artificial Intelligence and Machine Learning with the Physical Sciences.
The research was made possible by the usage of the SPRING workstation, provided by the company Mollificio Astigiano.
This work led to a collaboration with prof. Chris Rackauckas with the goal of testing and optimizing the NeuralPDE.jl library and other further development that are currently ongoing.

3.2.2 Contribution to the paper about the NeuralPDE library: Automating Physics-Informed Neural Networks (PINNs) with Error Approximations

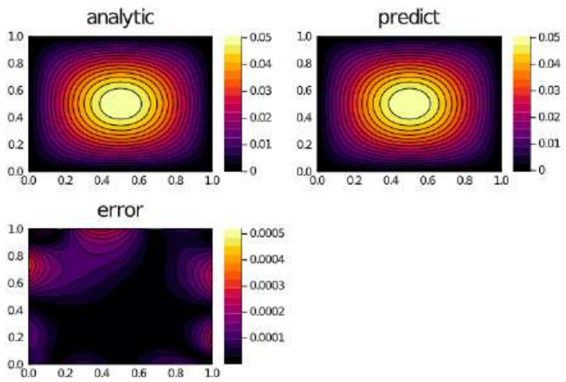
Pre-print: `https://arxiv.org/abs/2107.09443`
GitHub Code: `https://github.com/ChrisRackauckas/PINN_Quadrature`
Status: **Completed**
Participants: Francesco Calisto, Valerio Pagliarino, Simone Azeglio, Luca Bottero

We got in contact with Professor Christopher Rackauckas during the ProjectX competition and went on to collaborate with the MIT Julia Lab on a paper about the NeuralPDE.jl library, since we were one of the first research groups to test the Julia SciML ecosystem on a real world problem. Our contribution mainly consisted in the section devoted to quantitative comparisons between various training strategies and minimizers when solving different kinds of differential equations.

Physics-informed neural networks (PINNs) are an increasingly powerful way to solve partial differential equations, generate digital twins, and create neural surrogates of physical models.
NeuralPDE.jl is a PINNs library for the Julia language that uses the *Flux* backend for neural network training and the *ModelingToolkit.jl* tool for symbolic math problems formalization.

This approach is useful for tackling high dimensional problems, for ill-posed problems or in general when it is needed a non-conventional flexibility in constraining the solution. The physics-informed neural networks are very useful for training models of a physical system from a dataset constraining the solution to a *a-priori* physical knowledge.

We collaborated to a manuscript that explains in detail the inner workings of NeuralPDE.jl and shows how a formulation structured around numerical quadrature gives rise to new loss functions which allow for adaptivity towards bounded error tolerances.



In the paper, we describe the various ways one can use the tool, detailing mathematical techniques like using extended loss functions for parameter estimation and operator discovery, to help potential users adopt these PINN-based techniques into their workflow. We showcase how NeuralPDE uses a purely symbolic formulation so that all of the underlying training code is generated from an abstract formulation, and show how to make use of GPUs and solve systems of PDEs. Afterwards, we give a detailed performance analysis which showcases the trade-off between training techniques on a large set of PDEs. We end by focusing on a complex multiphysics example, the Doyle-Fuller-Newman (DFN) Model, and showcase how this PDE can be formulated and solved with NeuralPDE. Together, this manuscript is meant to be a detailed and approachable technical report to help potential users of the technique quickly get a sense of the real-world performance trade-offs and use cases of the PINN techniques [1].

As previously stated, we compared PINN models with different specifics based on their performance in terms of accuracy and run time. For example, we solved the 5 dimensional Hamilton-Jacobi-Bellman equation, which has the following form:

$$\frac{\partial u}{\partial t} + \nabla^2 u - \lambda ||\nabla u||^2 = 0 \qquad (2)$$

Here are a plot and a table showing the results, revealing that quadrature based methods outperform the standard *GridTraining* sampling technique on high dimensional problems:

[1] This sentence is a quote from the paper abstract, available on ArXiv.

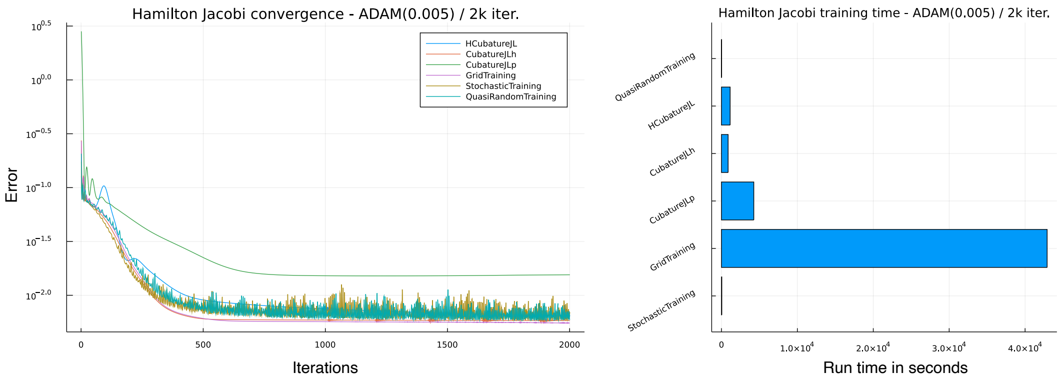


Figure 6: Convergence of the Hamilton Jacobi equation training using ADAM as minimizer.

Strategy	Minimum Error Value
HCubatureJL	5.92×10^{-3}
CubatureJLh	5.71×10^{-3}
CubatureJLp	1.51×10^{-2}
GridTraining	5.51×10^{-3}
StochasticTraining	5.90×10^{-3}
QuasiRandomTraining	5.85×10^{-3}

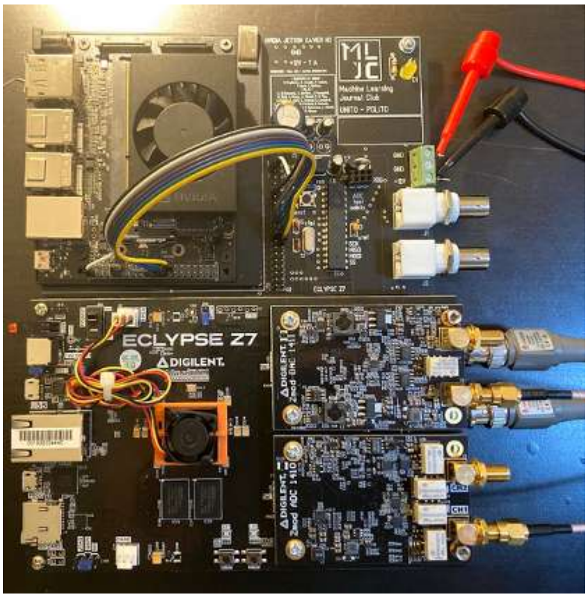
The SPRING workstation and the HPC4AI computational resources were necessary to train the examples needed for the comparisons.

3.2.3 Real Time Scientific Machine Learning Project

Status: Hardware completed, work in progress on model implementation and case studies
Participants: Francesco Calisto, Valerio Pagliarino, Simone Azeglio, Luca Bottero

After gaining an initial experience on the modeling of physical problems with architectures that allow the fusion of the knowledge coming from recorded data with the one coming from the laws of physics, we asked ourselves if it were possible to employ these models not only for dynamical system identification, but also for control purposes. In particular, we started asking ourselves if it were possible to run these machine learning-based models for system identification and automatic control with low latency, on edge-computing electronics. This goal partially overlapped with our interest in discovering new hardware platforms where running our model with some advantages, such as optimization, portability or inference speed. We are referring, in particular, to edge-computing GPUs, FPGAs and hybrid SoC+FPGAs chips. With this interest in mind, we started a new explorative project: "Real Time Scientific Machine Learning", with the main goal of introducing our group to the use of new technologies and methods.

The aim of this project is to investigate how some promising machine learning architectures can be implemented on specific embedded electronics (SoCs, GPU-accelerated SoCs, FPGAs) for real time applications, spacing from dynamical system identification, to control system, to signal processing. This project consists in a first phase that includes the development of a custom evaluation board embedding a SoC+FPGA subsystem, a GPU-Accelerated SOM (System on module) and some high-speed interfaces such as 14 bit 100 MSa/s ADCs and DACs.



The second phase consists in the in-depth study, implementation and qualifications of some architectures with particular attention to time-series analysis, physics-informed machine learning and reinforcement learning.
This activity is an ambitious long-term scientific project, considering that sometimes it will require to us to implement some machine learning models in hardware description language. It is possible thanks to the collaboration with NPO Torino that supplied some electronics boards and thanks to the computing power given by HPC4AI and the SPRING workstation for the model training and the HDL design and synthesis. A the end of 2021 the first phase has been successfully completed and we are starting the in-depth study of the architectures. We focused our attention on some specific models:

- Deployment of the *Mini-Rocket* [2] time series classifier. This architecture will be useful for the low-latency identification of trigger signals in a large variety of fields, with particular focus on biomedical signals coming from EEG and EMG. This allow our electronics to support the Brain-computer interfaces projects of the BCI group. Furthermore, this architecture can be used for other purposes such as fast signal identification in physics experiments. The figure 7 shows the block schematic of the setup used to qualify our model: a programmable signal generator is used to produce arbitrary signals that are used to train the real-time classifier running on the Zynq and reading one of the fast Digilent ADC input channels. An additional *Red Pitaya* programmable board can be used to provide custom and fast signals to our electronics while processing a possible output coming from the fast DAC channels of our board automatically. (Topic: Signal processing).

- Physics-Informed Neural Networks (PINNs) based on *NeuralPDE.jl* [3] for forecasting the evolution of dynamic systems with partial knowledge of physical laws by approximating the PDE solution with a neural network using boundary conditions coming from real-time sensor data. (Topic: Physics-Informed ML system identification and control in real-time).

[2] `https://arxiv.org/abs/2012.08791`
[3] `https://arxiv.org/abs/2107.09443`

- Physics-Informed reinforcement learning with various applications in physics and engineering.

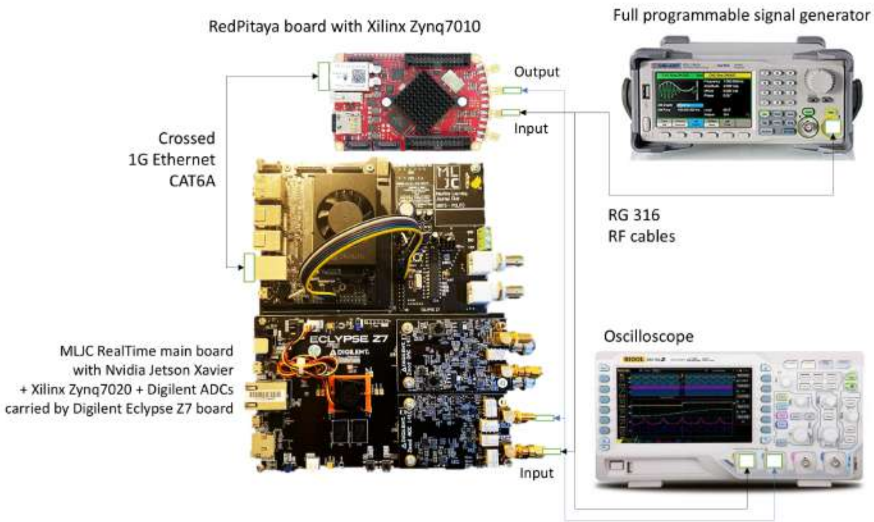


Figure 7: Block schematic of the setup used for training and testing the Mini-Rocket fast series classifier running on Xilinx Zynq.

While some of these architectures are currently under implementation or testing and will provide, in the future, powerful tools for innovative projects that requires interfacing with the hardware, we are already planning an ambitious project that will apply some of these techniques in the field of nuclear and medical physics. Nothing has already been published about about the *Real Time ML* project, being everything *work in progress*)

As Beatrice Villata joined our group, we are starting the preliminary evaluation phase of a project that will uses scientific machine learning techniques running with low latency (possibly deployable on real-time electronics) to make some intelligent control system to be used in the field of nuclear and medical physics, in particular radiotherapy, for adaptive beam control.

3.2.4 Ocean: Employing Physics Informed Neural Networks (PINNs) in Plastic Transport Problems

GitHub Code: `https://github.com/MachineLearningJournalClub/Ocean`
Status: Feasibility study completed, suspended, call for new participants open.
Participants: Andrea Semeraro, Mattia Falco, Francesca Bellucci, Simone Azeglio.

Andrea Semeraro, Mattia Falco, Francesca Bellucci and Simone Azeglio in this project started to apply Physics Informed Neural Networks for investigating problems related to ocean circulation. The aim is to devise techniques which can be helpful in tackling the problem of plastic debris in oceans. Currently, they are working on the Langmuir circulation problem with PINNs.

Selected Projects

Symmetries

3.3 Symmetries

- **Spokesperson**: Simone Azeglio
- **Participants**: Simone Azeglio, Arianna Di Bernardo, Simone Poetto
- **External collaborators**: Marco Nurisso, Luca Savant-Aira

Recently (December 2021), we have settled a group on deepening the role of Symmetries in the learning process, we are currently exploring several topics. For example:

- What could be the relationship between inductive biases in Geometric Deep Learning and Physics Informed Machine Learning models;
- Symmetry regularization as a biologically plausible form of regularization;
- On the ML side, investigate the role of disentangling representations - particular attention to Variational Autoencoders.
- Is ANNs' expressivity enough to capture the transformations occurring in different areas of our brains?

Related to the last one, we competed in the Brain-Score 2022 Competition, hosted at COSYNE (COmputational and SYstems NEuroscience) 2022, the most important conference worldwide in Computational and Systems Neuroscience.

3.3.1 Brain-Score 2022 Competition at COSYNE 2022

Status: Short paper accepted in the Brain-Score Workshop ; 3rd Place overall in the Brain-Score 2022 Competition
Brain-Score Workshop Paper: `https://arxiv.org/abs/2203.11910` **GitHub Code**: `https://github.com/sazio/brainscore2022`
Participants: Simone Azeglio, Simone Poetto, Marco Nurisso, Luca Savant-Aira

The aim of this project is to devise biologically plausible mechanisms for primate object recognition.

Computational models of vision have traditionally been developed in a bottom-up fashion, by hierarchically composing a series of straightforward operations - i.e. convolution and pooling - with the aim of emulating simple and complex cells in the visual cortex, resulting in the introduction of deep convolutional neural networks (CNNs). Nevertheless, data obtained with recent neuronal recording techniques support that the nature of the computations carried out in the ventral visual stream is not completely captured by current deep CNN models. To fill the gap between the ventral visual stream and deep models, several benchmarks have been designed and organized into the Brain-Score platform, granting a way to perform multi-layer (V1, V2, V4, IT) and behavioral comparisons between the two counterparts. In our work, we aim to shift the focus on architectures that take into account lateral recurrent connections, a ubiquitous feature of V1. Through recurrent connections, the input's long-range spatial dependencies can be captured in a local multi-step manner and, as introduced with Gated Recurrent CNNs (GRCNN), the unbounded expansion of the neurons' receptive fields can be modulated through the use of gates. In order to increase the robustness of our approach and the biological fidelity to specific layers of the ventral visual stream, we employ specific data augmentation techniques in line with several of the scoring benchmarks. Enforcing some form of invariance, through heuristics, was found to be beneficial for better neural predictivity.

As suggested by abundant physiological evidence, recurrent circuits are ubiquitous in the visual cortex and in several areas of the mammalian brain . In visual layers, the effect of lateral recurrent connections is believed to contribute to receptive fields adaptation. More generally, there are many indications of the capability of the brain to modulate the processing of visual signals on the basis of their context. Classical CNNs, despite showing many similarities with the ventral visual stream, lack the ability of context modulation. Trying to incorporate recurrent connections in a CNN architecture poses a problem because standard Recurrent Neural Networks (RNNs) are designed to process time varying sequences of inputs, while computer vision models deal with static inputs. To both circumvent this limit and try to design a more biologically plausible architecture, the basic idea is to introduce a form of recursion across neurons of the same layer. In this way it is possible to avoid the problem of static inputs, but still give to each neuron in a layer information about the activity of the surrounding neurons, allowing them to receive information about a larger part of the image. Similar solutions have been proposed in different forms by many authors, often with the aim of better simulating the visual system. It is worth noticing that the computational graphs of these models, when unfolded in time, look like pure feedforward hierarchies of operations enriched with a number of skip connections. This results in architectures that are very similar to residual networks. To achieve even better biological adherence, it is useful to add gates in between the recursive computations, as proposed by. In this way the neurons' receptive fields are explicitly modeled and modulated by the gates.

The introduction of gates is a distinctive feature of GRCNN: their role can be intuitively and qualitatively understood as an extra layer of computation resembling an attention mechanism. Gates are designed to give an output between 0 and 1 that multiplies (pointwise) the activations in the recurrent convolutional layer. This means that, during training, the set of weights associated with each gate evolves in such a way as to give the network the capability to notice which parts of the image are relevant. In more detail, a GRCNN is composed by a feedforward sequence of blocks called gated recurrent convolutional layers (GRCL). Every GRCL block computes a recursion on its inputs, and, between each recurrent operation, the gate system modulates the effective amount of forwarded information.
The equations describing the computations inside the GRCL are the following:

$$\begin{cases} x_0 = \mathcal{A}_0(u) & \\ g_t = \mathcal{B}_t(x_{t-1}) + \mathcal{C}(u), & t = 1, 2, \ldots, T \\ x_t = x_{t-1} + \sigma(g_t) \odot \mathcal{A}_t(x_{t-1}), & t = 1, 2, \ldots, T \end{cases} \tag{3}$$

where u is the input, $\mathcal{A}_t, \mathcal{B}_t, \mathcal{C}$ are convolutional operators with nonlinear activation functions (ReLU) and batch normalization, σ is the logistic sigmoid function and $\odot$ is the Hadamard

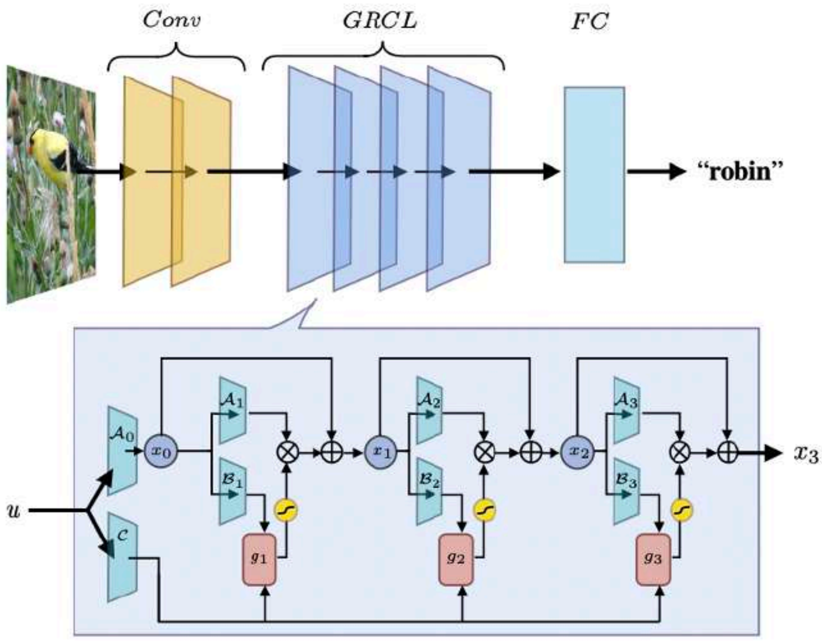


Figure 8: **Top** GRCNN architecture trained on the ImageNet dataset. It is composed by 2 Convolutional layers, 4 GRCL blocks and a Fully Connected layer which is employed for classification. **Bottom** Schematic representation of a GRCL block. Note that if we unfold the computational graph through time, we end up with something that is very similar to a ResNet50

product (elementwise multiplication operator). We emphasize the variables x_t and g_t which represent respectively the recurrent state variables and gate activations. The third equation clearly shows how the absence of gates would lead GRCL to be a standard ResNet$T$.

One of the main limitations in devising a deep model that predicts the neural activity of the ventral visual stream is the training procedure. Usually, vision models are trained on the ImageNet data set, which in its original formulation is related to a multiclass classification problem. On the other side, Brain-Score's benchmarks include several tasks which are not necessarily related to classification. In our work, we try to alleviate this problem by employing and designing augmentation and regularization strategies, inspired by some of the evaluation benchmarks. In particular, given their importance in the overall scoring, we decided to focus on behavior - more details can be found in -, V1 and V2 tasks based on Freeman & Ziemba benchmark.

Given the spirit of the behavioral benchmark, we took advantage of *CutMix*, an augmentation strategy that facilitates the recognition of different objects from partial views in a single image. More specifically, in CutMix, patches are cut and pasted among training images while, at the same time, the corresponding labels are linearly combined, resulting in: $\tilde{x} = \mathbf{M} \odot x_A + (\mathbf{1} - \mathbf{M}) \odot x_B$ and $\tilde{y} = \lambda y_A + (1 - \lambda) y_B$, where $\mathbf{M} \in \{0, 1\}^{W \times H}$ denotes a binary mask indicating where to drop out and fill in from two images, $\mathbf{1}$ is a binary mask filled with ones and λ is sampled from the uniform distribution $U(0, 1)$.

In parallel, to enhance the robustness of learned representations, we employed *AugMix*, which allows our model to explore the semantically equivalent input space around an image. Briefly, AugMix consists in combining simple augmentation operations - e.g. translation, rotation, shear - together with a consistency loss. Augmentations are sampled stochastically and concatenated while a consistent embedding around the input image is enforced by using the Jensen-Shannon divergence as a loss.

Lastly, we introduced a regularization term, with the idea that behavioral traits cannot be fully described in terms of a scalar metric - e.g. accuracy - but need to be conceived in terms of higher-order descriptors such as reconstructing a confusion matrix, in order to force a network to fail in the same way as a primate would. We took as a reference human performances on 11 superclasses (i.e. groups of classes) of ImageNet, projected down the predictions of our model - 1000 dimensional vector - onto the 11 superclasses and imposed an additional cross entropy loss on this term, to quantify the distance between our model and human performances.

To improve scores on V1 and V2 benchmarks, we opted for an augmentation technique largely inspired by Freeman & Ziemba benchmark. Given that V2 neurons are particularly sensible to textures while V1 neurons respond in a similar way to a texture and its phase randomized counterpart, we considered a texture dataset composed of images coming from different open dataset, and we extended it by generating each texture's phase randomized counterpart, named *noise* [4]. After that, we fine-tuned our model's first GRCL block - roughly corresponding to V1 - by freezing all the other layers and randomly blending input images separately with textures and noise. Later on, we fine-tuned the second GRCL block in the same way, but by only blending input images with textures. In this way we tried to emulate how specific layers in the visual cortex respond to different stimuli, and we got improvements in both V1 and V2 with respect to the baseline.

Introducing receptive fields modulation through a gated recurrent mechanism is beneficial in terms of neural predictivity in the visual cortex. The baseline GRCNN model shows promising results in several benchmarks. As shown in Table 1, by introducing the previously mentioned regularization and augmentation techniques, scores improve on specific benchmarks, as well as on average.

In the current state, further work is needed to get a better sense of such preliminary results [5]. In this regard, we have implemented several variants of the baseline model, including an integration of the VOneBlock as a substitute of the first convolutional block in the GRCNN. With more computational power we are planning to train this combined architecture from scratch. Ultimately, another interesting perspective for further exploration is related to how CNNs learn textures and not shapes, which is ultimately analogous to our augmentation strategy for V1 and V2.

[4] We open-sourced generated Textures and Noise dataset at https://github.com/sazio/TexturesNoiseDataset
[5] We open-sourced the code for our experiments at https://github.com/sazio/brainscore2022

Name	Average	V1	V2	V4	IT	Behavior
GRCNN55 Behavior	.463	.509	.303	.482	.467	.554
GRCNN55	.462	.525	.306	.481	.479	.520
GRCNN109	.461	.520	.328	.475	.464	.521
GRCNN55 V1-V2	.458	.535	.314	.486	.481	.473
ResNet50	.427	.497	.264	.465	.475	.432
AlexNet	.424	.508	.353	.443	.447	.370

Table 1: Brainscore benchmarks results

3.3.2 ICLR - Geomstats Computational Geometry & Topology Challenge 2022

Status: Ongoing, deadline April 4th 2022.
Participants: Simone Azeglio, Arianna Di Bernardo

The purpose of this challenge is to foster reproducible research in geometric (deep) learning, by crowdsourcing the open-source implementation of learning algorithms on manifolds. Participants are asked to contribute code for a published/unpublished algorithm, following Scikit-Learn/Geomstats' or PyTorch's APIs and computational primitives, benchmark it, and demonstrate its use in real-world scenarios.
Each submission takes the form of a Jupyter Notebook leveraging the coding infrastructure and building blocks from the package Geomstats. The participants submit their Jupyter Notebook via Pull Requests (PR) to the official (https://github.com/geomstats/challenge-iclr-2022) GitHub repository.
In addition to the challenge's prizes, participants will have the opportunity to co-author a white paper summarizing the findings of the competition.

In our case, we decided to focus on Riemannian geometry and topological data analysis applications on neural recordings, in particular, electroencephalogram (EEG) data. The idea is to exploit the intrinsic structure of the covariance matrix, which is a Positive Semi-Definite matrix, thus living on the SPD manifold. Covariance matrices are of particular interest in brain computer interfaces and in general whenever spatially localized sensors are gathering data. By employing geomstats, we designed a Riemannian SVM, which, coupled with topological features, extracted from EEG signals, lead to better results in terms of accuracy, in a human multi-class classification problem of a motor task.

3.3.3 Symmetries Workshop - Bernstein Conference 2022

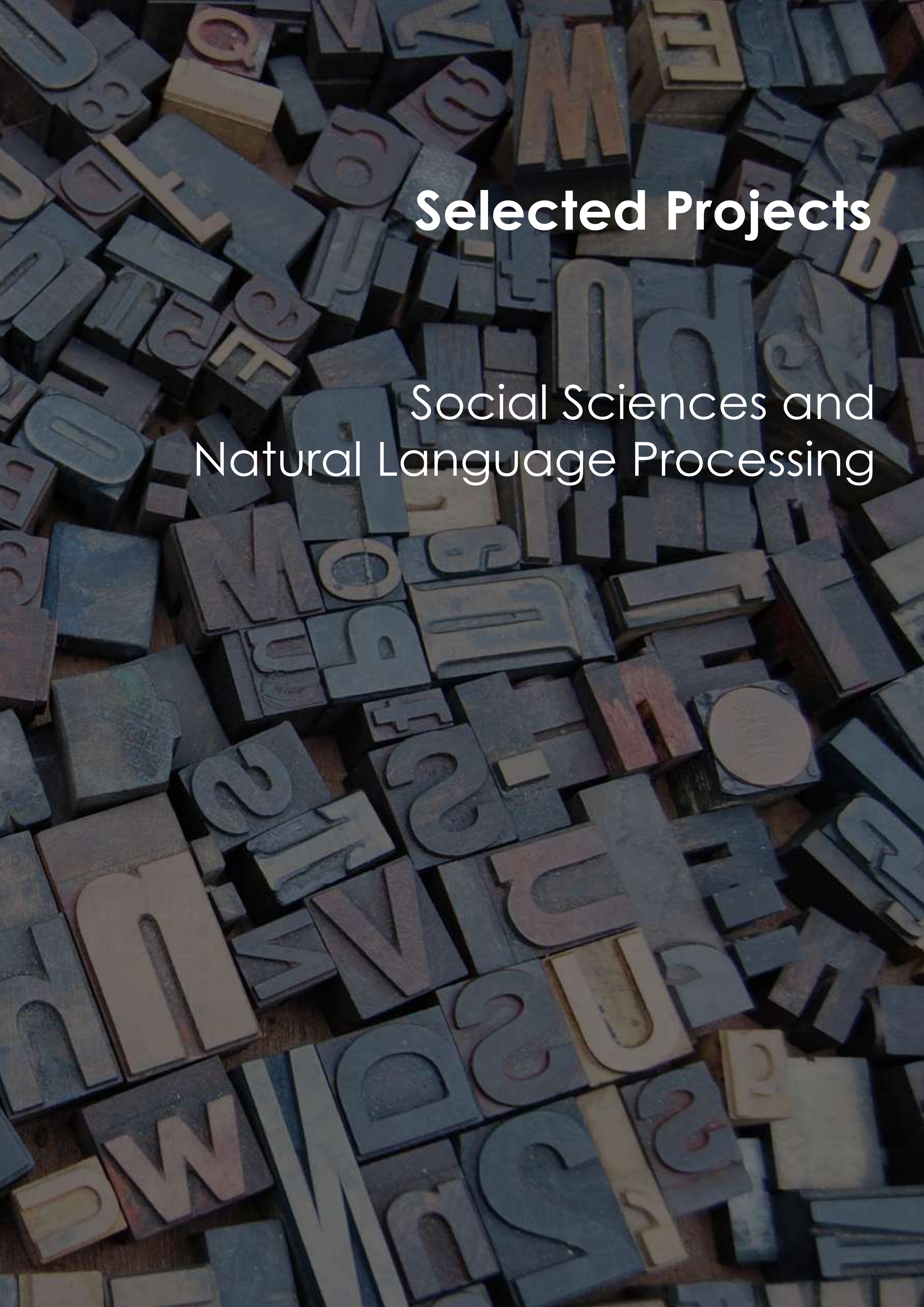
Status: Submitted, waiting for results.
Participants: Simone Azeglio, Arianna Di Bernardo

The Bernstein conference, organized by the Bernstein Network, based in Berlin, has established itself as one of the most renowned computational neuroscience conferences worldwide. In this context, we submitted a two-days workshop proposal focused on the main topic of this group: the role of symmetries and invariances in neural representations.

Workshop Title: Symmetry, Invariance and Neural Representations

Abstract. Why does symmetry encoding matter for neuroscience? There exists an intimate relationship between how natural phenomena evolve in time and how we represent the world by "measuring" it with our senses. On the one hand, the best mathematical model of the world we possess, the physical laws, can be characterised on the basis of invariants, conserved quantities (Noether's Theorem). On the other hand, in order to perceive and thus interact with the external environment, we need to create robust neural representations from the 'data' collected through sensory processing. It is therefore inevitable that neural representations, as well as the organisation and dynamics of neuronal circuits, are affected by the organisational properties dictated by physics. Awareness of this fact and incorporating it in the definition of computational models for brain function can allow for more robust learning and provide better generalisation properties. This workshop proposal aims to unify under a common framework, theories and models that consider invariant representations of the brain function.

Relevance. This proposal is one of the first attempts towards creating a common framework to investigate invariant neural representations through the lens of symmetries. With this workshop, we want to show how the concept of symmetry shapes different neural mechanisms ranging from vision, audition, olfaction, locomotion, memory. Defining a unified perspective for synthesizing robust representations is a relevant and exciting opportunity for designing better computational models of brain function.



Selected Projects

Social Sciences and Natural Language Processing

3.4 Natural Language Processing and Social Sciences

- **Spokesperson**: Marina Rizzi
- **Participants**: Marina Rizzi, Simone Azeglio, Micol Olocco, Luca Bottero, Pio Raffaele Fina, Alessio Borriero
- **External collaborators**: Luca Pezzini, Alberto Guffanti, Weronika Sójka, Zofia Piętka-Danilewicz, Piotr Pranke

This group was created with the aim of learning the most advanced tools of text analysis and natural language processing for analyzing text data, that can be used in the social and economic sciences and in the humanities. The group started its activities by organizing some open-access online meetings, with a didactic aim. Group members have been involved in several online competition and community weeks regarding NLP (i.e. The CommonLit Readability Challenge, The HugginFace's Flax/JAX community week). In the Academic Year 2021/2022, the main activities of this group were to organize divulgative meeting at the Toolbox Coworking, discussing various topics regarding machine learning, text analysis, scraping, and in general, issues regarding practical data analysis and computer science.
All the material produced by the NLP and Social Sciences group can be found in our GitHub Repos:

- for Learning NLP: `https://github.com/MachineLearningJournalClub/LearningNLP`
- for NLP and Scraping at Toolbox: `https://github.com/MachineLearningJournalClub/NLP-and-Scraping-at-Toolbox`.

3.4.1 Natural Language Processing and Social Sciences - Teaching Meetings

Status: Completed
Code: The materials of our meetings can be found on Github: `https://github.com/MachineLearningJournalClub/LearningNLP`.
Participants: Simone Azeglio, Marina Rizzi, Luca Bottero, Micol Olocco, Alessio Borriero

During the 2020/2021 Academic Year, we organized some online meetings to discuss the recent text analysis techniques that are used in the analysis of text data. We had over 70 registrations for our meetings, and in the majority of the meetings we had more than 40 participants. We talked about topics regarding text preprocessing, logistic regression for text classification, topic modeling, ethical bias in the use of machine learning models, models used for competitions like the "CommonLit Readability Challenge" on Kaggle, and other state-of-the-art models for what concern text analysis and natural language processing.

- **Tutorial 1**: Sentiment Analysis, Vectorizing words & Explainability Methods.
- **Tutorial 2**: Bias & Fairness in NLP, Gender Framing and Topic Modeling through Latent Dirichlet Allocation;
- **Tutorial 3**: Exploratory Data Analysis for the CommonLit Readability Dataset, Feature Extraction from Pretrained Word2Vec, Dimensionality reduction and visualization with UMAP and Data Augmentation for Word2Vec

- **Tutorial 4**: Global vectors for word representations (GloVe); Fasttext, skipgrams vs CBOWs; Bias in Word Embedding with WEFE and possible causes

---

3.4.2 CommonLit Readability Challenge on Kaggle

Status: Completed, model submitted and evaluated
Code available here `https://tinyurl.com/CommonLitReadabilityChallenge`
Participants: Simone Azeglio, Marina Rizzi, Luca Bottero, Micol Olocco, Alessio Borriero

Can machine learning identify the appropriate reading level of a passage of text, and help inspire learning?
Currently, most educational texts are matched to readers using traditional readability methods or commercially available formulas. However, each has its issues. Tools like *Flesch-Kincaid Grade Level* are based on weak proxies of text decoding (i.e., characters or syllables per word) and syntactic complexity (i.e., number or words per sentence). As a result, they lack construct and theoretical validity. At the same time, commercially available formulas, such as *Lexile*, can be cost-prohibitive, lack suitable validation studies, and suffer from transparency issues when the formula's features aren't publicly available.
CommonLit, Inc., is a nonprofit education technology organization serving over 20 million teachers and students with free digital reading and writing lessons for grades 3-12. Together with Georgia State University, an R1 public research university in Atlanta, they are challenging Kagglers to improve readability rating methods.
In this competition, we designed a machine learning model based on sentence transformers, coupled with topological features coming from word-embeddings to rate the complexity of reading passages for grade 3-12 classroom use. Our model have been successfully submitted and evaluated 748th over 3638 different models.

3.4.3 The HuggingFace x Flax/JAX community week 2021

Status: project completed, code available here `https://tinyurl.com/flax-sentence-embeddings`
Participants: Simone Azeglio

HuggingFace partnered-up with Google's Flax, JAX, and Cloud teams to organize a new community week from July 7th to July 14th. During this week, we learnt how to effectively use JAX/Flax for Natural Language Processing (NLP) and Computer Vision (CV). Free access to a TPUv3-8 VM has kindly been provided by the Google Cloud team.
The goal of the HuggingFace x JAX/Flax community week is to make compute-intensive NLP and CV projects (like pre-training BERT, GPT2, CLIP, ViT) practicable for a wider audience of engineers and researchers.
Participants could propose ideas for an interesting NLP and/or CV project. Teams of 2 to 5 were formed around the most promising and interesting projects. In our case we designed a large sentence transformer model (+1 billion parameters) in Nils Reimers' group.

3.4.4 NLP and Scraping at Toolbox

Status: Ongoing, monthly meetings
Participants: Marina Rizzi, Micol Olocco, Luca Bottero, Pio Raffaele Fina, Alberto Guffanti, Luca Pezzini.

During the 2021/2022 Academic Year, we organized some open meetings at Toolbox, a co-working based in Turin. The starting plan was to focus on topics related to Natural Language Processing and Scraping, but we afterwards decided to enlarge our range of topics covered, organizing meetings about topics regarding machine learning in general. The topics we covered during our meetings comprehend scraping of web pages, how to extract information in a systematic way from PDF, Word and PowerPoint documents, which bias (and in general, which ethical bias) are present in machine learning models, and in models of natural language processing, model of continual learning, and so on. The GitHub material can be found at the following link: `https://github.com/MachineLearningJournalClub/NLP-and-Scraping-at-Toolbox`.

3.4.5 Latin writing styles analysis with Machine Learning: New approach to old questions

Status: Completed
Pre-print: Available on arXiv `https://arxiv.org/abs/2109.00601`
GitHub Code: `https://github.com/machinelearningjournalclub/thus-torun`
Participants: Arianna Di Bernardo, Pietro Sillano, Simone Poetto, Beatrice Villata, Weronika Sójka, Zofia Piętka-Danilewicz, Piotr Pranke

We have been invited, from the University of Torún, to spend three weeks in the History department. Thanks to this opportunity, four of us spent this time in Poland, having the chance to learn from a new environment and apply our knowledge in a new field.
In particular, the project focused on the analysis of Middle Ages texts. We considered a specific construction of literature composed in Latin and indicated the probability patterns of familiar sources of specific narrative texts. We carried out the task consisting of the practical use of those concepts and observations to create a tool for analyzing narrative texts using *similies* for whole phrases. Based on open-source databases, the probability of the origin of specific text sources has been calculated. We focused on creating search tools resources that could enable detailed search throughout the text. The main objectives of this study took into account finding similarities between sentences and between documents. We also performed machine learning algorithms on chosen texts to calculate their specific features, for instance, authors, centuries, or appearing words. One extra objective was to apply machine learning algorithms to recognize sources of anonymous texts with a certain percentage. Consideration of Natural Language Processing tools allowed us to transform textual objects into numerical ones and then apply machine learning algorithms to extract information from the data set.
This work resulted in writing an article, which was then published on ArXiv.

Competitions, events & partners

4 International competitions

4.1 ProjectX 2021 by University of Toronto UofT AI

ProjectX is an international competition organized by the University of Toronto. It lasted for five months, during which students from the bachelor's studies designed and performed a scientific project connected to machine learning. This year's topic was *Health* and our Turin group was the only Italian one involved in this competition, among several American ones. We worked on a project about immunology. Our idea was centered along with the design of a tool to cluster in an unsupervised way antigens starting from the information collected from mass spectrometry experiments. The abstract of our work is reported as follows:

Abstract

Despite the worldwide effort to gain a deeper comprehension of the molecular mechanism that characterizes cancer, its blueprint is still elusive. In the last decade, results in the context of cancer immunology helped researchers to identify a set of biological objects, namely antigens, which now represent a promising direction. Mass spectrometry is the typical investigation involved in collecting molecules. This method and modern techniques allow a limited but satisfactory throughout on peptide knowledge. However, the diversity coming out from such experiments is still poorly characterized. In the paper, UnChAnTies (Unsupervised characterization of antigen communities) is introduced. This tool is a computational platform that enables unsupervised clustering and identifies critical variables characterizing each community of antigens. The goal of the pipeline is to identify relevant structures through unsupervised feature extraction with a model-free approach. Additionally, shed light, that takes part in the interplay between various kinds of cancer through their antigen are determined.

4.2 EO Dashboard 2021

From June 23-29, 4300+ participants from around the world joined NASA (National Aeronautics and Space Administration), ESA (European Space Agency), and JAXA (Japan Aerospace Exploration Agency) for the all-virtual, global Earth Observation Dashboard Hackathon. Among the 10 challenges related to the COVID-19 pandemic, our team (Luca Bottero, Micol Olocco, Luca Pezzini, Andrea Semeraro, Beatrice Villata) focussed on ”A comparative Analysis” topic where the pandemic's economic impacts in urban areas for the USA, Asia, and Europe were to be evaluated sing data from the Earth Observing Dashboard (the “EO Dashboard” for short).

Summary

The COVID-19 pandemic has had different impacts in different regions of the world. The EO Dashboard Hackathon `https://www.eodashboardhackathon.org/` allowed people from all over the world to tackle the problem of analyzing such differences from several points of view. During the competition, we faced the comparative analysis challenge `https://tinyurl.com/5y7v5ezb`. This repository collects all the files used to perform such analysis, presenting some novel approaches aimed to improve the current EODashboard.

Background

2020: under the pandemic wave, the world slowed down. Industries and transports were heavily impacted and, beyond the effects on private lives and social interactions, the pandemic had repercussions on the environment. The question leading us was how much of these changes we could see in data from satellites. We focused on variations in greenhouse gases (SO2, NO2, O3, CO and CH4) and differences in these values across various world countries. To this end, we performed a time-oriented analysis on time series and a complementary geography-oriented analysis, based on unsupervised feature extracting, to highlight differences between countries.

Data Analysis

Being part of the Machine Learning Journal Club, we approached the problem from our knowledge of Data Analysis and Machine Learning. For the time-oriented analysis, we employed the Prophet algorithm, implemented by Facebook's Core Data Science team to do forecasting prediction `https://facebook.github.io/prophet/`. As a complementary route, the geography-oriented analysis aimed to quantify the differences in behaviors of a group of countries in Europe, Asia, and America during the pandemic: to this end, we used a method that performs dimensionality reduction on the time series of the given gases (embedding). That is achieved using a Neural Network Autoencoder, a perceptron that tries to create an output that is equal to its input. We trained the NN and we got the output on the intermediate layer of it (in our specific model it is 9-dimensional). Roughly speaking, we "compressed" the information from the time series in a single vector, namely the internal representation in the autoencoder. We then combined the different vectors given by every time series for a given country (one for every gas) by joinin

Objectives

Our ultimate goal is to create a feasible application to the EODashboard to quantify the socio-economic impacts that the pandemic had on different countries. We used disparate variables from the Copernicus Sentinel 5 as proxies; in particular, we employed SO2, NO2, O3, CO and CH4. More variables can be added to improve the work.

GitHub repository: `https://github.com/MachineLearningJournalClub/EOChallenge`
Further details here: `https://tinyurl.com/2s3uxnmx`

5 Events

5.1 Participation to AAAI

We have participated to the 2021 AAAI Spring Symposium on Combining Artificial Intelligence and Machine Learning with Physics Sciences, where we presented our work on Physics Informed Machine Learning applied to Wildfire Spread Prediction in the form of a short talk. It was a highly formative experience, where we have had the chance to keep up to date with the latest researches in Machine Learning, Scientific Machine Learning in particular. Here you can find the proceedings of the conference: `http://ceur-ws.org/Vol-2964/`.

Figure 9: AAAI - Machine Learning and the Physical Sciences Symposium 2021

5.2 Participation to COSYNE

We took part in the COSYNE 2022 meeting, in particular by presenting a talk during the Brain-Score Workshop on computational model of object detection that are able to predict neuronal activity of the ventral visual stream. During this 6 days congress in Lisbon and Cascais (Portugal), we had the chance to be part of the vibrant and international community of computational neuroscientists and better grasp the role of ML in Neuroscience.

Figure 10: COSYNE 2022

5.3 Collaboration with Toolbox Coworking

Toolbox Coworking is a coworking space created near Torino Porta Nuova station, in Via Agostino da Montefeltro, 2. It consists of an interactive and hybrid space, where freelancers, companies and start-ups work every day to strengthen and innovate their business. In the Academic Year 2021/2022, we started a collaboration with Toolbox Coworking, they provide us with rooms and infrastructures to organize monthly meetings on topics related to machine learning and natural language processing techniques. These meetings are completely open-access (both in the Toolbox community and outside), and they are advertised on Eventbrite. The material produced for the meeting at Toolbox is available here: `https://github.com/MachineLearningJournalClub/NLP-and-Scraping-at-Toolbox`. More information about Toolbox Coworking can be found here: `https://toolboxcoworking.com/en/explore-toolbox`.

5.4 Sharper researchers' night

The Machine Learning Journal Club participated in the SHARPER – European Researchers' Night in Turin on September 24th 2021, organized by the University of Turin and Politecnico of Turin, in the beautiful frame of the Castello del Valentino. The title of the stand was "MACHINE LEARNING E L'INTELLIGENZA ARTIFICIALE DEL FUTURO". At our stand we presented projects spacing from Brain Computer Interfaces to Scientific Machine Learning. Our poster and other multimedia contents can be found on our web page: `https://www.mljc.it/sharper-european-researchers-night-2021-turin/`.

Figure 11: Our stand at the SHARPER European Researchers' Night 2021

6 Partners

The Machine Learning Journal Club is an officially affiliated student association of the University of Turin (`https://tinyurl.com/mljc-albo-unito`) with official headquarter at the Department of Physics, Via Pietro Giuria 1. We acknowledge the University of Turin for their essential support. Our activities are possible only thanks to the valuable collaboration with some important institutions and companies: in 2021 MLJC collaborated with the followings.

6.1 NPO Torino

`https://www.nposervices.com/`

NPO Torino is a company part of the Ricoh group that offers services in the field of ICT and Digital Transformations. NPO Torino is a fundamental partner of MLJC for hardware technology, electronics, on-premise infrastructures and application of ML techniques to industrial case studies. MLJC deeply thanks NPO Torino for their contribution and their valuable support during projects development, hackathons and all our activities. We are very grateful to Francesco Dipietromaria, Romualdo Delmirani and Massimo Altamore for their continual support and trust in our association.

Figure 12: Some members of the Machine Learning Journal Club receiving from NPO electronics components for research projects.

6.2 HPC4AI and Computer Science Department of UNITO

`https://hpc4ai.unito.it/`

Figure 13: Some members of the Machine Learning Journal Club visiting HPC4AI in October 2021

HPC4AI is the Turin's High-Performance Centre for Artificial Intelligence (coord. Prof. Marco Aldinucci). The first goal of HPC4AI is to establish a large and modern laboratory to co-design with industries and SMEs research and technology transfer projects. HPC4AI has been co-funded by Regione Piemonte via EU POR-FESR 2014-2020 with 4.5M€ and started its activities at beginning of 2019.

HPC4AI@UNITO is a research infrastructure which aims at designing, developing and operating experimental, innovative applications and services at the convergence between High-Performance Computing and Artificial Intelligence. HPC4AI@UNITO is part of the federated cloud infrastructures organised in 4 data-centres funded with 4.5M€by Piedmont Region, the University of Torino and the Polytechnic University of Torino under the INFRA-P 2018 call.

HPC4AI provides Machine Learning Journal Club with computing power, GPU-enabled virtual machines and cloud technology. MLJC deeply thanks HPC4AI for their important contributions. In particular, we would like to show our gratitude to Marco Beccuti, Marco Aldinucci and Claudio Mattutino.

6.3 Toolbox Coworking

Toolbox Coworking is a coworking space created near Torino Porta Nuova station, in Via Agostino da Montefeltro, 2. It consists of an interactive and hybrid space, where freelancers, companies and start-ups work every day to strengthen and innovate their business. In the Academic Year 2021/2022, we started a collaboration with Toolbox Coworking, where they provide us rooms and infrastructures to organize monthly meetings on topics related to machine learning and natural language processing techniques.

Figure 14: Some pictures taken during our meetings at Toolbox Coworking

The material produced for the meeting at Toolbox is available here: `https://github.com/MachineLearningJournalClub/NLP-and-Scraping-at-Toolbox`. More information about Toolbox Coworking can be found here: `https://toolboxcoworking.com/en/explore-toolbox`.

6.4 Pompei Student Lab

`https://www.pompeilab.it/`

Pompei Student Lab is an initiative of Don Luca Peyron that provides young people spaces and resources for studying and for other impactful activities. Pompei Student Lab provided MLJC with ample spaces for co-working and to host some of our servers.

Figure 15: MLJC members and external collaborators at the Pompei Student Lab

6.5 Mollificio Astigiano

Mollificio Astigiano is a leading designer and producer of custom springs for industry, precision mechanics and agriculture. Their philosophy meets state-of-art production using innovative technologies with a particular attention to the society. The MLJC appreciate the particular passion of Mollificio's board and staff for science, innovation and their attention for young people and for the educational world. Mollificio Astigiano provided MLJC with a workstation optimised for parallel computing in 2020.

7 Funding

The Machine Learning Journal Club has been funded by the University of Turin for the Academic year 2021-2022 with € 2600.


READY
FOR
THE
CHALLENGE
?
Time
to grow