EPJ Web of Conferences **302**, 17005 (2024) https://doi.org/10.1051/epjconf/202430217005
SNA + MC 2024

Research Directions on AI and Nuclear

Daniela Cancila[1*], *Geoffrey* Daniel[2], *Jean-Baptiste* Sirven[3], *Zakaria* Chihani[1], *Fabian* Chersi[1], and *Regis* Vinciguerra[1]

[1]Université Paris-Saclay, CEA, List, F-91120 Palaiseau, France
[2]Université Paris-Saclay, CEA, DES, F-91120 Palaiseau, France
[3]Université Paris-Saclay, CEA, Service de Physico-Chimie, F-91120 Gif-sur-Yvette, France

Abstract. The development of applications and systems for the nuclear domain involves the interplay of many different disciplines and is, therefore, particularly complex. Additionally, these systems and their innovations have to be compliant with strict international regulations and recommendations. The scientific and industrial communities have been studying, developing and applying advanced Artificial Intelligence (AI) techniques and tools in several (non-nuclear) application domains. Their encouraging results have pushed the nuclear community to pay increasing attention to the field of AI. Among the expected benefits of AI is the simplification of complex procedures, the reduction in the execution of time-consuming operations, the increase of safety levels, and the reduction in the overall cost. At the French Atomic Energy Commission (CEA), we have identified and have started to address several open questions, such as: where in the nuclear domain can AI-based techniques be implemented in the most productive way? What do the nuclear standards and recommendations say about its use? Can we identify some core challenges and issues common to multiple areas of the nuclear domain? In this paper we provide a first analysis and answers to the above questions and we conclude by emphasizing some cross-domain high priority challenges.

1 Introduction

In recent years, we have witnessed a large increase in the use of Artificial Intelligence (AI) in many fields of applied computer science. It has become more and more evident that AI-based methods can efficiently solve complex specific problems, where other techniques struggle.

Without any doubt, the nuclear industry is among the most conservative domains of application, and is subject to the most restrictive safety regulations. Let us consider IEC 60880 [1] and IEC 61513 [2] as a typifying example of these regulations. IEC 60880 specifies the safety objectives and requirements for critical software, whilst IEC 61513 concerns safety objectives and requirements for Instrumentation and Control (I&C) systems and equipment to perform safety functions in a Nuclear Power Plant (NPP). Both regulations require that all measures, including accuracy, safety margins, uncertainty, response time and execution time (the list is non-exhaustive), must be validated and proven.

* Corresponding author: daniela.cancila@cea.fr

At the same time, digitization is affecting many domains of application. For example, advanced simulation techniques employ “digital twins”, i.e. the ability to integrate multi-physics parameters at different layers of abstraction - thus creating a digital representation of a physical object or event, and relating it to the physical world.

With full confidence, we can assert that a convergence between digital twins and AI will happen in the near future. For example, predictive maintenance of a component or a system exploits advanced analysis based on huge amounts of data. In this regard, AI represents a natural option to provide inferential conclusions.

In our view the domain of application of nuclear systems cannot be exempt from this digital revolution. To our knowledge, for several years, the scientific and industrial communities in the nuclear fields have been increasingly paying attention to digital twins and AI-based techniques. Some tools are already available as open-source, such as MiLady (Machine Learning Dynamics), which aims “to improve the accuracy and predictive power of atomistic simulations […] with a reasonable computational cost” [3]. For applications in nuclear industry, AI is seen as a very powerful tool, which could be used with increased awareness of both its potential and limitations.

From a pragmatic point of view, the use of AI for nuclear applications requires transversal collaboration of two disjoint communities: the one of AI software engineers and the one of nuclear engineers interested in AI. To close that gap, in January 2023, we have held a first seminar between two units of the French Alternative Energies and Atomic Energy Commission (CEA): List [4] and ISAS [5]. This article originates from that seminar. Note that the analysis of the fields and the challenges identified here are not intended to be exhaustive, but constitute a first in a series of reports on this subject.

The objective of this paper is twofold. First, we identify two disciplines, spectroscopy and robotics both related to the inspection activity, in which AI could provide advanced solutions. Second, we ask ourselves if it is possible to abstract from the analysed fields by identifying core and cross-fields challenges. Our reflections consider the aforementioned applications as a whole and conclude the paper by emphasizing some cross-domain high priority challenges.

2 State of the art and regulations

Of the extensive literature on AI-related issues on the nuclear domain, we here discuss existing ecosystems and their work, that are closer to our activity. A specific state of the art is added in sections 2 and 3.

Figure 1 represents a simplification of the international standardization and regulation ecosystem (ISO/IEC JTC1, IEC/SC 45A, IAEA).

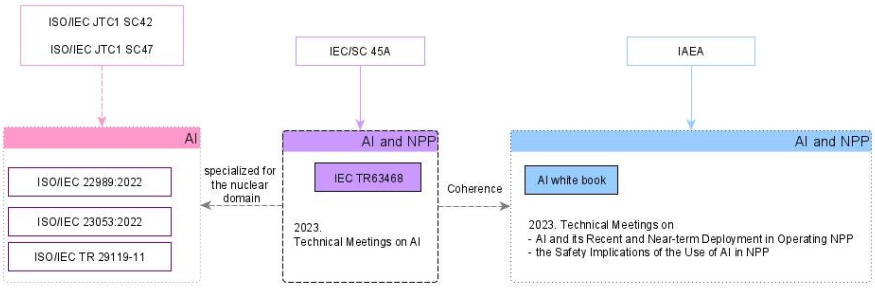

Fig. 1. International standardization and regulation ecosystem

The International Atomic Energy Agency (IAEA) is particularly active in the investigation of AI in nuclear domain. In 2022, the book "Artificial Intelligence for Accelerating Nuclear Applications, Science and Technology" [6] identified some disciplines for which AI could provide some benefits. Chapters 7-10 and 12 span from data to nuclear physics, from fusion to nuclear power (which highlight for example the benefit of AI for automation, design, optimization, data analytics, prediction and prognostics, and insights extraction) to verification (including gamma spectroscopy and robotics). Without any surprise, AI is identified as a useful tool when large amounts of data need to be analysed.

The ISO/IEC Joint Technical Committee (JTC1) addresses the international standards on Information technology. In the last years, the subcommittees on "Artificial Intelligence" (SC42) and on "Software and systems engineering" (SC7) have devoted an important effort to publish standards [7] [8] and technical reports [9]. These works are characterized by being domain-specific independent and aim to harmonize AI over different application domains, by proving general rules and the same terminology to adopt.

Finally, the IEC SC45A "Instrumentation, control and electrical power systems of nuclear facilities" is devoting some effort – and not without a live debate – to understand whether and how to standardize AI for the nuclear domain. In this context, a recent publication is IEC/TR 63468:2023 "nuclear facilities – instrumentation and control, and electrical power systems – artificial intelligence applications" [10] which is based on [7] [8] [9] (list non exhaustive). Among the main highlights of this technical report, we underline the classification of the main AI approaches and a clear identification of the international standards to be followed if a new domain-specific standard for AI in the nuclear domain should be considered necessary.

Some reflections and challenges are shared by the above groups, IAEA, ISO/IEC JTC 1/SC42 and SC7, and IEC/45A (the following list is non-exhaustive):

- The difficulty of Verification & Validation for AI and the proposal of some solutions
- Ensuring cyber-security of AI systems.

3 Structure of the article

Figure 2 represents the structure of this paper. Given the limited space, we only discuss two topics: one in nuclear and one in computer science (having applications in the nuclear domain). "Trustworthy AI" is the core challenge, which has arisen from our studies.

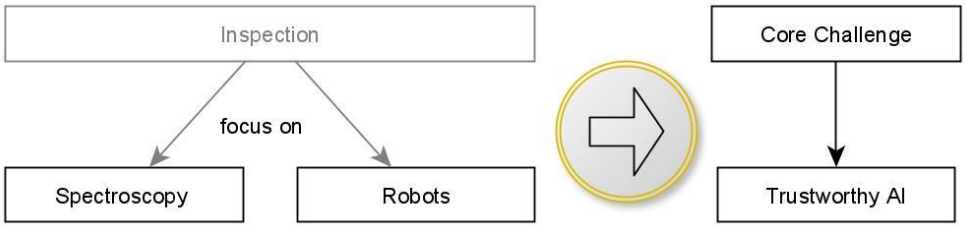

Fig. 2. Studied domains and identified core challenge

Spectroscopy and robotics are related to the inspection activity. Independently of the techniques and tools adopted by the inspection and the technology of the NPP itself (e.g. Gen-IV, EPR, SMR), the activities covered by the inspection will always be conducted in NPP; in other words, they are inherent to the NPP itself and essential for a safe development and maintenance of the nuclear energy. In this regard, innovation concerns techniques, methods and tools used for the inspection activities. Techniques based on digital twins, Machine Learning (ML) and AI have proven to be extremely beneficial to other application

domains. Among their distinguishing features, there is the ability to ingest massive volumes of heterogeneous data (which NPP do produce) and to use them to infer correlations between data and events inductively. It is therefore worth studying whether and how these innovative methods can be beneficial to inspection activities of NPPs.

3.1 Artificial intelligence and spectroscopy for materials analysis in nuclear and industrial contexts

The chemical analysis of materials or components is at the heart of many nuclear processes, whether to ensure certain safety functions, to control the correct operation of a process, or to check the conformity of a product. In terms of on-line analysis, optical techniques make it possible to carry out measurements at a distance on samples of all kinds and in a few seconds to a few minutes. They are therefore well suited for real-time analysis in difficult environments, particularly in radioactive ones [11]. Furthermore, they also have applications in laboratories, for example for the rapid characterization of samples or the in situ monitoring of processes at the pilot R&D scale.

These techniques are based on optical radiation spectroscopy, and include scattering (Raman spectrometry), fluorescence (laser-induced fluorescence - LIF), absorption (near-infrared spectroscopy - NIRS, UV-visible spectrometry, cavity ringdown spectroscopy - CRDS), or laser ablation (laser-induced breakdown spectroscopy - LIBS). Thanks to their short analysis time, they can quickly generate a large volume of data. Moreover, as analytical techniques without sample preparation, the signal they deliver can be strongly influenced by the variability of the laser-matter interaction related to the physical properties of the samples. Finally, the coupling of phenomena, e.g. in LIBS the sampling by laser ablation and the resulting plasma emission, adds further complexity to the interpretation of the spectra.

These reasons have motivated the development of multivariate approaches for several decades to efficiently exploit spectral data, extract maximum information and optimize analytical performance. In the particular field of spectroscopy, chemometrics is the discipline that brings together these approaches. Historically, it covers linear methods such as principal component analysis, independent component analysis, partial least squares regression and many other techniques, but over time it has also integrated non-linear approaches such as support vector machines or neural networks (NN) [12]. With the development of AI, and especially with the last ten years of deep learning, the processing of spectroscopic data for chemical analysis has taken a new turn. Many possibilities are provided by these modern approaches, with potentially increased performances compared to the usual techniques.

In analytical spectrochemistry, tools based on Deep Neural Networks (DNNs), Convolutional Neural Networks, Auto-Encoders or Long Short-Term Memories are now used for many applications. These include signal denoising [13] or instrumental artefact removal [14], classification of compounds or samples [15], reconstruction of spectra within mixtures [16], supervised analysis for the prediction of physico-chemical properties or concentrations (quantitative analysis) [17, 17], calibration transfer [18], processing of hyperspectral imaging data [19] or data fusion [20]. The issue of trustworthiness of predictions has also been recently addressed [21].

In the nuclear field, the challenges for artificial intelligence in spectroscopy are of several kinds. The first one is the very large volume of training data required by many deep learning algorithms, which is often incompatible with what can be obtained experimentally, particularly when the implementation constraints are strong (radiation protection, quantities of material available, etc.). Simulation of training data (data augmentation) is a promising avenue to overcome this limitation. From an analytical point of view, the reliability of predictions is also a crucial issue and requires rigorous model validation steps. The quantification of uncertainties is also essential. Finally, the explainability of the models is

necessary to guarantee a good control of the algorithms and the way in which information used to make a decision is extracted from the spectra.

Beside these issues, which will have to be addressed in the future, the possibilities that can be foreseen in terms of AI in analytical spectroscopy concern, for example, the correction of matrix effects in quantitative analysis, the exploitation of time-resolved spectral data, and the detection of anomalies, the development of tools to ensure the robustness of predictions, high-throughput and high-resolution spectral imaging, approaches combining modelling and AI, the use of large spectral databases shared within experimental platforms, and the downsizing of instrumentation.

3.2 Autonomous mobile robots

Robots are means to assist in the inspection of a NPP site. The benefit in their use with AI has been pointed out for example in [6]. In this section, we focus on mobile autonomous robots. We anticipate that their use can be even more widespread in a SMR (small modular reactor) technology, aiding to reduce the maintenance cost of a NPP under operation.

Constraints arising from energy consumption and, more generally, the resource limitations that characterize all robot's embedded devices pose serious challenges. For example, in order to decrease the possibility of putting the robot's mission (i.e. inspection) at risk, the execution of control algorithms (e.g. navigation, obstacle detection, perception of the environment) should be calibrated to have a low resource footprint. In particular, it is worth stating the obvious: the smaller the robot, the more challenging it becomes to embed all the required control-and-command devices within it.

Another crucial and complex characteristic is the robustness and precision of the robot's self-localization. CEA List has developed specific AI-based methods, for example in sensor fusion, perception and neural fields, having a very low resource footprint [22].

Additionally, CEA List is developing a framework called Aidge [23] for optimizing and deploying DNNs on low-power architectures, such as those that are implemented in autonomous mobiles robots. In particular, this framework is simple and fast to use, without requiring advanced knowledge in deep learning or hardware. It allows rapid and effortless DNN applications generation and porting to a great variety of hardware.

4 Core Challenge: Trustworthy Artificial Intelligence

If we consider the two areas as a whole, we observe that two main aspects return as a leitmotif for a possible deployment of these tools and techniques in the nuclear industrial context.

The first concerns the quantification of the uncertainty, which ought to be associated with each result of the analysis. This requirement is demanded by the nuclear safety standards. Moreover, this quantification plays an important role because it allows nuclear engineers to know the level of trustworthiness of the achieved results.

The second leitmotif addresses the robustness of machine-learning models. These two observations require a detailed investigation and are the scope of the two following subsections.

4.1 Uncertainty quantification in Artificial Intelligence

Training an ML model usually corresponds to approximating an unknown function that represents the data. In supervised ML approaches, this function is a mapping between the inputs and expected outputs of the model. This approximation leads to unavoidable errors in

the prediction that can be more or less important depending on the training data and the test data. In the context of trustworthy AI, it is necessary to address this problem and recent state-of-the-art approaches focus on the paradigm of uncertainty quantification.

The main objective is to associate an estimation of the uncertainty to the prediction of a ML model. This uncertainty can come from intrinsic variability inherent to the data (measurement errors, statistical fluctuations, lack of information due to hidden variables), called aleatoric uncertainty, and from the lack of knowledge (limited training samples, extrapolation) or the lack of complexity of the model, called epistemic uncertainty.

The main challenge is the reliable evaluation of these uncertainties, especially in the domain of Deep Learning algorithms. Because of the high number of parameters in DNNs, often associated to the high dimensionality of the data, conventional statistical or Bayesian methods are not tractable. Over the past decade, several approaches have been proposed to obtain estimations of these uncertainties (e.g. Quantile Regression, Bayesian approaches with Variational Inference, Monte-Carlo Markov Chains, Monte-Carlo Dropout) [24]. These methods provide useful elements for uncertainty quantification; however, none of them has emerged to be consensual. It is thus necessary to understand in detail the benefits and the limitations of these methods, and to always be cautious with the interpretation.

In nuclear applications, reliable uncertainty quantification methods are clearly important when decisions are based on AI predictions. We can use a NN for a thermal-hydraulics simulation code: the NN provides a rapid answer compared to the simulation code. However, by using uncertainty quantification, the NN indicates whether its prediction is reliable, and thus suggest whether it would be preferable to use the simulation code instead.

Nevertheless, because of the ongoing developments of these methods, another challenge is the validation of the uncertainties provided by the models. Validation methodologies have been proposed based on coverage plots [25] and out-of-distribution data detections [26], and recalibration techniques can be applied to improve the uncertainty estimation [27]. However, they rely mainly on empirical evaluations by using test data, and the question still remains whether AI models are used in abnormal conditions that have not been anticipated. All of these problems must be addressed to make ML models acceptable for deployment in critical applications, even if they can outperform other conventional methods in several domains.

4.2 Robustness of Artificial Intelligence models

Recent studies have shown that conventional DNNs present weaknesses concerning robustness [28]: a NN can easily be misled by small perturbations, known as adversarial attacks, in the input data that completely modify the model's prediction. Famous examples show images of cats that are normally classified as cats by a NN, but are classified as dogs after applying small perturbations on pixels that are invisible to human eyes.

These perturbations can be computed by exploiting the differentiability of the NN to find the optimal small modification to apply to the data in order to disturb the NN. These attacks are called "white-box" attacks [29] and they require knowing the structure and the parameters of the NN. On the contrary, it has been shown that "black-box" attacks [30] can also be applied and they require only the possibility to utilize the model for predictions.

It is important to highlight that this weakness comes from the use of NN on high-dimensional data, such as images. In high-dimensional spaces, the higher degrees of freedom give more possibilities to find a direction of perturbation that misleads a model [31].

The main challenge is to defend the models against these attacks. Adversarial training has been proposed [32] and consists of feeding the NN with adversarial examples during the learning phase. Other studies focus on the regularity of the models [33]. There is indeed a strong connection between regularity and robustness: a model that is regular in every direction of the space is less susceptible to be misled by a perturbation. Certification methods

[34] try to give boundaries for the predictions of the NN, assuming a controlled magnitude of the perturbation. However, a trade-off tends to emerge between the application of these methods and the performance of the models. Uncertainties quantification can also address this issue, by detecting possible attacks through a reduction of the epistemic level of confidence. These characteristics of AI models must be considered for critical applications, such as nuclear systems.

Whether these attacks are intentional, which involves cyber-security challenges, or they are due to random noise in measurements [35], methods of defense are necessary to enhance the robustness of the models, especially to obtain trustworthy AI.

Finally, other types of attacks must be considered. For instance, while deploying large scale AI, the models can be manipulated by corrupting the training database, called "poisoning attacks" [36], in order to introduce malicious biases in the predictions. The models could be hacked to expose privacy data that would have been used in the training database, by using model inversion attacks [37].

4.3 Towards qualification tools

Software qualification (also known as software certification for example in the avionic application domain) is a complex process, which demands (non-exhaustive list): (a) methods and qualified tools, often based on a mathematical model, for the design and the development of the software itself; and (b) methods and qualified tools to verify and test the adequacy to safety-related standards.

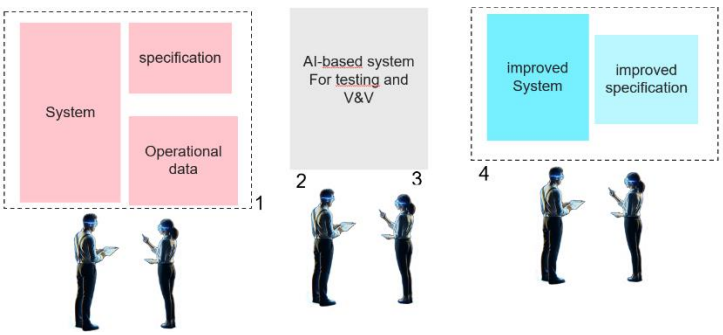

Fig. 3. Safety engineers analyse the system (1) and ask to a qualified AI-tool to test the system (2). the generated documentation (3) provides an improved system (4) under the engineer's supervision (4). Image partially created with Copilot and optimized via Dall-E-3.

This situation is even more complex if the software utilizes AI-based algorithms, and/or if the tools used for software qualification exploit AI. To the best of our knowledge, these topics are still under investigation and an international consensus is not fully established, despite the recent standards (e.g. [7]) and a technical report [10].

A qualification process for AI inherits a considerable number of aspects from the "traditional" software qualification process. Usually, it starts with a comprehensive review of the AI-based system's design, architecture, and functionality, which typically relies on carefully crafted key performance indicators (KPIs) [38], such as accuracy, precision, and robustness, as well as identifying any potential risks and vulnerabilities.

Depending on the adopted AI method and properties to prove, different approaches can be chosen for verifying the software and the behavior explicability. For example, techniques based on supervised ML can be adopted to predict the behavior of a reactor, and the "metamorphic testing" approach [39] among the means to verify the supervised ML.

5 Conclusions and Outlook

We can with full confidence assert that a convergence between AI and nuclear applications will happen in the near future. At the beginning of this work, we identified the following questions: where can AI-based techniques be implemented in the most productive way for the nuclear domain? What do the nuclear standards and recommendations say about its use? Can we identify core challenges and issues common to multiple areas of the nuclear domain?

The first question covers a vast area of possible applications and disciplines, and would deserve an entire paper on its own. These applications are characterized for example by different levels of sensitivity, and where and when they are used (e.g. in the design & development, testing, verification of a system, or embedded in an alarm system or in a robot).

We think that all nuclear disciplines that are faced with processing of a very large amount of data could obtain a benefit of AI-based tools. In the less sensitive ones, tools exploiting new AI-generating technologies, such as nuclear-adapted versions of GPT [40], can allow an important acceleration e.g. in the analysis of documents and traceability. However, the more the area in question concerns safety-related data, the more it will be necessary to have control over the quality of the results. For this purpose, there are several strategies that we could explore, such as a parallel execution of AI-based with traditional techniques (diversification), or, even better, a combination with traditional techniques.

Although, to the best of our knowledge, AI is not currently allowed in a NPP, international organizations, such as IAEA and IEC 45A, are intensifying their effort in such an investigation, see e.g. [6] [10]. The underlying idea is to achieve a consensus in the industrial community and the production of international regulations. Our outlook is to actively contribute to the work, led by IAEA and IEC 45A.

Finally, AI trustworthiness is the achievement of several independent studies we have conducted at CEA [4] [5]. It has been identified as a core challenge common to different areas of the nuclear domain. During our work, we identified a strictly related emerging challenge: cybersecurity. In the near future, we foresee a greater attention towards cybersecurity issues, due to the digitization in the nuclear domain, the adoption of machine-learning- and neural-network-based tools, especially together with the development of SMR technologies because, in given configurations, operators could remotely control the nuclear power plant. To our knowledge, IAEA and IEC 45A have also highlighted the same challenge to the safe deployment of AI in a NPP.

6 Acknowledgements

We would like to thank: Patrick Landais, former Haut-Commissaire of CEA, for the constructive discussions and the identification of the cybersecurity challenge; Alexander Bounouh, head of the CEA List, and Erwan Adam, head of the “Service de Génie Logiciel pour la Simulation” for supporting the cross-activity on “AI and Nuclear” between CEA List and CEA ISAS; Lorène Allano and Andrea Zoia for a first review of this paper; Samuel Evain for the notes, and all the participants of the “AI and Nuclear” seminar and the colleagues that have supported this initiative.

7 References

1. IEC 60880, Nuclear power plants – Instrumentation and control systems important to safety – Software aspects for computer-based systems performing category A functions (2006)

2. IEC 61513, Nuclear power plants – Instrumentation and control important to safety – General requirements for systems (2011)
3. M. C. Marinica, et al., MiLaDy - Machine Lerning Dynamics (CEA, Saclay, 2015-2019) https://ai-atoms.github.io/milady-docs/ (visited January 2024)
4. CEA LIST, https://list.cea.fr/en/ (visited January 2024)
5. CEA DES ISAS, https://www.linkedin.com/company/cea-isas/ (visited January 2024)
6. IAEA, Artificial Intelligence for Accelerating Nuclear Applications, Science and Technology (2022)
7. ISO/IEC 22989:2022, Information technology – Artificial intelligence (2022)
8. ISO/IEC 23053:2022, Framework for Artificial Intelligence (AI) Systems Using Machine Learning (ML) concepts and terminology (2022)
9. ISO/IEC TR 29119-11, Software and systems engineering – Software testing – Part 11: Guidelines on the testing of AI-based systems (2020)
10. IEC/TR 63468:2023, Nuclear facilities – instrumentation and control, and electrical power systems – artificial intelligence applications (2023)
11. D. Doizi et al., Optical On Line Techniques for Nuclear Applications., Proceedings of ANIMMA 2011 (2011)
12. S. Brown, R. Tauler, B. Walczak, Comprehensive Chemometrics. Chemical and Biochemical Data Analysis (Second Ed.), Elsevier (2020)
13. T. Yoon et al., Deep learning-based denoising for fast time-resolved flame emission spectroscopy in high-pressure combustion environment, Combustion and Flame, **248** (2023)
14. S. P. Kyathanahally et. al., Deep learning approaches for detection and removal of ghosting artifacts in MR spectroscopy, Magnetic Resonance in Medicine, **80**, pp. 851-863 (2018)
15. C. S. Ho et al., Rapid identification of pathogenic bacteria using Raman spectroscopy and deep learning, Nat Commun, **10** (2019)
16. Z. Zhao et al., ConInceDeep: A novel deep learning method for component identification of mixture based on Raman spectroscopy, Chemometrics and Intelligent Laboratory Systems, **234** (2023)
17. W. Ng et al., The influence of training sample size on the accuracy of deep learning models for the prediction of soil properties with near-infrared spectroscopy data, SOIL, **6**, p. 565–578 (2020)
18. P. Mishra et al., Deep calibration transfer: Transferring deep learning models between infrared spectroscopy instruments, Infrared Physics & Technology, **117** (2021)
19. R. Finotello et al., HyperPCA: A powerful tool to extract elemental maps from noisy data obtained in LIBS mapping of materials, Spectrochim. Acta Part B., **192** (2022)
20. H. Leng et al., Raman spectroscopy and FTIR spectroscopy fusion technology combined with deep learning: A novel cancer prediction method, Spectrochim. Acta Part A, **285** (2023)
21. R. Finotello, D. L'Hermite, C. Quéré, B. Rouge, M. M. Tamaazousti, J.-B. Sirven, Trustworthiness of Laser-Induced Breakdown Spectroscopy Predictions via Simulation-based Synthetic Data Augmentation and Multitask Learning, EPJ Web of Conferences (2023)
22. J. Coulin, R. Guillemard, V. Gay-Bellile, C. Joly, Tightly-coupled magneto-visual-inertial fusion for long term localization in indoor environment, IEEE Robotics and Automation Letters (2021)
23. F. Chersi, et al., AIDGE: A Framework for Deep Neural Network Development, Training and Deployment on the Edge, Proceeding of the European Conference for Edge AI (2023)

24. J. Gawlikowski, C. R. N. Tassi et al., A survey of uncertainty in deep neural networks, arXiv preprint arXiv:2107.03342 (2021)
25. T. Pearce, A. Brintrup et al., High-quality prediction intervals for deep learning: A distribution-free, ensembled approach, International Conference on Machine Learning (PMLR), p. 4075–4084 (2018)
26. K. Lee, H. Lee et al., Training confidence-calibrated classifiers for detecting out-of-distribution samples, International Conference on Learning Representations (2018)
27. Y. Romano, E. Patterson, E. Candes, Conformalized quantile regression, in Advances in neural information processing systems, 32 (2019)
28. C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, et al., Intriguing properties of neural networks., arXiv:1312.6199 (2013)
29. X. Ma, Y. Niu, L. Gu, Y. Wang, et al., Understanding adversarial attacks on deep learning based medical image analysis systems, Pattern Recognition (2021)
30. C. Guo, J. Gardner, Y. You, A. G. Wilson, K. Weinberger, Simple black-box adversarial attacks., in In International Conference on Machine Learning (2019)
31. I. Goodfellow, , J. Shlens, et al., Explaining and harnessing adversarial examples., arXiv preprint arXiv:1412.6572 (2014)
32. A. Madry, A. Makelov, et al., Towards deep learning models resistant to adversarial attacks., arXiv preprint arXiv:1706.06083 (2017)
33. H. Gouk, E. Frank, et al., Regularisation of neural networks by enforcing lipschitz continuity, Machine Learning, **110**, pp. 393-416 (2021)
34. J. Cohen, E. Rosenfeld, et al., Certified adversarial robustness via randomized smoothing, Proceedings of the 36th International Conference on Machine Learning Research, **97** (2019)
35. Z. Chaouai, et al., Application of adversarial learning for identification of radionuclides in gamma-ray spectra, Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment, **1033** (2022)
36. B. Nelson, M. Barreno, et al., Exploiting machine learning to subvert your spam filter., LEET, **8**, n. 1-2, pp. 16-17 (2008)
37. M. Fredrikson, S. Jha, et al., Model inversion attacks that exploit confidence information and basic countermeasures, In Proc. The 22nd ACM SIGSAC conference on computer and communications security, pp. 1322-1333 (2015)
38. J. Mattioli, H. Sohier, et al., An overview of key trustworthiness attributes and KPIs for trusted ML-based systems engineering, In AITA: AI Trustworthiness Assesment at AAAI (2023)
39. T. Chen, Metamorphic testing: A simple method for alleviating the test oracle problem, In Proc. of the 10th International Workshop on Automation of Software Test. AST '15, IEEE Press (2015)
40. OpenAI, GPT-4 https://openai.com/product/gpt-4, (visited Mai 2023)