Research Challenges in Trustworthy Artificial Intelligence and Computing for Health: The Case of the PRE-ACT project

Foivos Charalampakos[§*], Thomas Tsouparopoulos[§*], Yiannis Papageorgiou[§*], Guido Bologna[†], André Panisson[‡], Alan Perotti[‡] and Iordanis Koutsopoulos[§]
[§]Athens University of Economics and Business, Athens, Greece
Email: {phoebuschar, tsouparop20, gpapageorgiou, jordan}@aueb.gr
[†]University of Applied Sciences and Arts of Western Switzerland
Email: guido.bologna@hesge.ch
[‡]CENTAI, Torino, Italy
Email: {andre.panisson, alan.perotti}@centai.eu

***Abstract*—The PRE-ACT project is a newly launched Horizon Europe project that aims to use Artificial Intelligence (AI) towards predicting the risk of side effects of radiotherapy treatment for breast cancer patients. In this paper, we outline four main threads pertaining to AI and computing that are part of the project's research agenda, namely: (*i*) Explainable AI techniques to make the risk prediction interpretable for the patient and the clinician; (*ii*) Fair AI techniques to identify and explain potential biases in clinical decision support systems; (*iii*) Training of AI models from distributed data through Federated Learning algorithms to ensure data privacy; (*iv*) Mobile applications to provide the patients and clinicians with an interface for the side effect risk prediction. For each of these directions, we provide an overview of the state-of-the-art, with emphasis on techniques that are more relevant for the project. Collectively, these four threads can be seen as enforcing *Trustworthy* AI and pave the way to transparent and responsible AI systems that are adopted by end-users and may thus unleash the full potential of AI.**

I. Introduction

Artificial Intelligence (AI) applications are rapidly permeating and transforming many areas of society and have already widespread use in the analysis of biological and medical data. The medical science is actively seeking opportunities to apply AI towards automating and optimizing notoriously difficult predictive tasks so as to provide tangible benefits to society. Examples include personalized healthcare, patient monitoring, drug development, advanced prognosis, among others.

The PRE-ACT project[1], under the full name "Prediction of Radiotherapy side effects using explainable AI for patient communication and treatment modification" is a Horizon Europe project that aims to apply AI to predict the risk of side effects following radiotherapy treatment for breast cancer patients. The project aims to deliver a framework, grounded on solid and novel AI concepts, for prediction of radiotherapy side effects and subsequently utilize it to inform about optimal treatment. The project's prime focus is on arm lymphoedema, since it is one of the most disabling side effects in the commonest cancer type, with comparatively high long-term survival. Other side effects are acute skin toxicity, late breast toxicity (e.g., breast atrophy), and cardiac toxicity. The project consortium includes researchers in the UK, Greece, Netherlands, France, Italy and Switzerland, with the combined expertise in computing, AI, radiation oncology, medical physics, genetics, psychology and health economics that is necessary to tackle this problem.

The project will leverage data from three multi-centre patient European cohorts to train AI models for risk prediction of the aforementioned side effects. Data include patient medical records such as comorbidities, anatomy, demographics, as well as treatment data, radiotherapy dose distribution data, Computerized Tomography (CT) scans, auto-contouring of critical organs in CT scans, and genetic data. The project will use these data to train AI models. A communication package that will emerge through a systematic participatory co-design methodology with patients and physicians, will ensure that predictions from the AI model will be presented in a meaningful, explainable manner to patients and clinicians to inform their joint decision-making regarding the choice of radiation treatment option.

The impact of explainability of the AI model will be assessed in a clinical trial that comprises two arms, namely two disjoint subsets of recruited patients. In the first arm, the personalised risk prediction will be communicated to physicians and patients, while in the second arm, it will not. Subsequently, in the patient followup period, the impact of explainability will be assessed through the rate of occurrence of arm lymphoedema, and through other quality of life indicators for patients. The ultimate goal is to demonstrate that risk communication in an explainable manner improves patient quality of life. In this paper, we provide an overview of a number of AI and computing technologies, we outline key points and motivate their use in the context of the project. These technologies are:

- Explainable AI models that demystify the hidden black-box AI prediction models through easy-to-understand explanations;
- Decentralized training of AI models using advanced Federated Learning algorithms;
- Fairness considerations in the AI algorithms that aim at uncovering and explaining potential biases in data.

[*]Equal contribution
[1]https://preact-horizoneurope.eu/

- Mobile and web applications that emanate from a co-design approach with patients and physicians and aim at presenting the risk prediction to them.

II. Explainable Artificial Intelligence

The evolution in the field of AI has led to the predominance of several contemporary and complex methods such as Neural Networks (NNs) when AI is applied to decision-making processes. In the health domain decisions have a large impact on individuals, yet modern AI methods provide a limited understanding of their predictions [1]. Due to the criticality of such domains, model understanding is essential so that humans trust the model's predictions. EXplainable Artificial Intelligence (XAI) aims to provide explainable predictions, leading to insights that can prove useful to various stakeholders such as patients, doctors, and regulatory authorities [1].

XAI methods can be arranged to several taxonomies according to different criteria such as the scope of the explanation, i.e., whether it concerns explanation of a single instance x regarding a single patient (local explainability) or explanation of the whole model behavior for all the health data it has encountered (global explainability). Local methods are further categorized based on the explainability principles that each algorithm is based on, as well as whether the algorithm is model-agnostic (i.e., it does not require access to the model architecture) or model-specific. In this section, we discuss two well-known families of explainability methods, namely feature attribution and propositional rules.

A. Feature Attribution

Feature attribution (FA) methods are explanation techniques that assign significance scores to features based on certain criteria. A feature's score captures its contribution to the predicted value of an instance. For example, given a single patient's data, the model predicts a certain probability for the side-effect of lymphoedema, and a feature attribution method can be used to highlight how critical each feature was to the model's prediction. One well-known method which can be used for FA, LIME (Local Interpretable Model-agnostic Explanations) [2], yields a local explanation by training an interpretable model on a new sampled dataset consisting of perturbed samples of the original dataset and the corresponding predictions of the "black box" model. Each new sample is also weighted by the proximity of the sampled instances to the instance of interest x. The features' importance is then acquired by an interpretable (surrogate) model, e.g., the weights of a linear regression model.

LIME's output is an inherently interpretable model g, acquired by minimizing the loss L (e.g., MSE) which measures how close the prediction of g is to the prediction of the original model f, while model complexity $\Omega(g)$ is kept low (for example, we prefer fewer features included). Finally, the proximity measure π_x defines the weighting scheme for the neighbors.

$$explanation(x) = \underset{g \in G}{\operatorname{argmin}} L(f, g, \pi_x) + \Omega(g) \qquad (1)$$

Feature attribution methods are a very promising subfield of XAI but their dependence on perturbations make them fragile and prone to adversarial attacks [3]. Methods to make them more robust as well as ways of combining them with concepts like Federated Learning (FL) where data access is not trivial, could constitute directions for future research.

B. Rule Extraction

Many early AI applications relied on expert systems that performed reasoning using propositional rules. As the present state of an expert system can always be described by the numerous inferences derived from the beginning state, these systems were intrinsically explainable. Consequently, converting the classifications of the NN models into propositional rules is a natural way to describe them.

A conclusion and one or more rule antecedents are both included in propositional rules. A rule antecedent A_i is formally defined as $a_i < t_i$, or $a_i \geq t_i$, where a_i is an input variable and t_i is a constant. Then, a propositional rule with k antecedents is: "if A_1 and ... and A_k then *Conclusion*". "Conclusion" is a class in the context of data classification. Golea demonstrated that explainability of Multi-Layer Perceptrons (MLPs) by means of propositional rules is NP-hard [4]. The number of rules and the number of antecedents per rule are two criteria that are frequently used to define the complexity of the extracted rules. Rulesets with less complexity are preferable, because they are easier to understand at first glance.

If it is possible to arrange data in tables, that is, in rows and columns where each piece of information is always placed in the same position, then the data is said to be tabular. Data represented by images is not tabular, because the same object could appear in multiple locations. Deep NNs are not clearly superior to well-known machine learning models, such as MLPs, Support Vector Machines (SVMs), ensembles of models, etc., for tabular data. Many scholars tried to use propositional rules to explain the knowledge that is included in MLPs and SVMs. A thorough survey of rule extraction methods to explain MLP responses was provided by Andrews et al. [5]. Furthermore, a review of numerous explainability strategies was presented for SVMs [6].

Whether propositional rules are ordered or not is a key aspect to consider. Ordered rules are given as:

if tests1 on antecedents are true then ...,
else if tests2 on antecedents are true then ...,
...,
else ...

The word "else" is absent in unordered rules. A sample can therefore activate more than one rule. Long ordered rulesets are challenging to understand because they may contain a large number of implicit antecedents, particularly those negated by "else if" (belonging to previous rules). As all rule antecedents are explicitly presented, unordered rulesets typically present more rules and antecedents than ordered ones, making them more transparent. As all antecedents are explicitly stated, each rule in an unordered ruleset reflects a distinct body of knowledge that can be analyzed separately. One would attempt to fully comprehend the significance of each rule in relation to the data domain while dealing with a large number of unordered rules. Acquiring the big picture could take some time.

There are nodes and edges in a binary decision tree (DT), which is a recursive structure. A predicate with respect to an

attribute is represented by each node. Based on its value, the classification route for a sample proceeds to either the left or right branch until it reaches a terminal node. A propositional rule is defined by each route leading from the root to a terminal node. DTs are therefore always comprehensible.

Model ensembles frequently offer greater accuracy than a single model. Several ensemble learning techniques, including bagging [7] and boosting [8], were proposed. They have been used with NNs and DTs. Yet, even DTs that can be simply translated into propositional rules, when coupled in an ensemble lose their interpretability. Two major techniques are used with ensembles of DTs to produce propositional rules. The first makes an effort to make DTs more diverse in order to decrease the number of DTs in an ensemble. As a result, with a small number of trees, all of the rules that each tree produces are considered. Examples of methods for diversity optimization are described in [9]. With the second group of methods, the main strategy is to remove as many rules as possible.

Few studies were achieved on NN ensembles for rule extraction. To produce unordered propositional rules from ensembles, one of the authors suggested DIMLP ensembles [10], [11]. Specifically, by pinpointing the exact location of axis-parallel discriminative hyperplanes, propositional rules were produced. The REFNE algorithm was introduced by Zhou et al. [12] (Rule Extraction from Neural Network Ensemble). A trained ensemble creates new samples in REFNE before extracting propositional rules. Additionally, attributes are discretized, and specific fidelity evaluation procedures are used. Finally, there were only three possible antecedents for rules. Johansson created rules from ensembles of 20 neural networks using the genetic programming method [13]. Here, the optimization problem of rule extraction from ensembles was seen as a trade-off between accuracy and comprehensibility. Finally, for a small ensemble of MLPs, Hara and Hayashi presented a rule extraction technique [14].

III. Fairness in Clinical Decision Support Systems

The increasing availability of Electronic Health Records (EHR) and the rapidly growing predictive power of AI models have contributed to both the advancement of research and the creation of business opportunities for deploying clinical Decision Support Systems (DSS) in healthcare facilities [15], [16]. With this growth, however, comes a concern that these models may learn undesired spurious correlations during the training process, with the resulting decisions being polluted by unintended biases. While there is growing interest in the AI community to address biases and fairness-related issues [17], [18], many of these efforts are still in the early stages of development. Quantitative and systematic auditing of real-world datasets and AI models is a nascent field, and there is a need for interdisciplinary approaches to define, investigate, and provide guidelines for tackling these issues.

The definition of *fairness* is a complex and multi-faceted concept, and its meaning can differ depending on the context and the perspectives of the partners involved [19]. In the context of AI and ML, fairness can refer to a variety of properties such as equal treatment, equal opportunity, and equal outcomes. Questions of fairness are often approached as mathematical problems where researchers and practitioners tend to adopt a quantitative perspective, focusing on developing models that meet specific criteria, such as equal allocation, representation, or error rates. This is often done by framing the task of model development as a constrained optimization problem, where fairness constraints are incorporated into the optimization process. The specific constraints used in this process may be informed by laws, social sciences, and philosophical perspectives. By using a quantitative approach, AI practitioners aim to optimize models for fairness, balancing it with other objectives such as accuracy or efficiency. In the context of clinical decision support systems, fairness may be defined as providing equal treatment to all patients regardless of their protected characteristics, such as gender or ethnicity.

Clinical ML models are often trained in heterogeneous contexts, possibly with data collected from different populations. When designed without special constraints regarding fairness, such models are likely to exhibit biases and non-fairness problems. Bias can arise from various sources, such as the data used to train the model, the choice of features and algorithms, or the distribution of the data across different groups. Therefore, it is crucial to evaluate the fairness of a model regardless of how it was developed. This evaluation can be performed using fairness metrics and by examining the model's predictions for different groups and demographic variables. The results of this evaluation can inform decisions about the deployment and usage of the model, and can also be used to identify areas for improvement and future work.

Seminal works [20], [21] have shown that increasing awareness about the possible fairness issues in a ML model leads to increased transparency. By thoroughly evaluating the fairness of a model and explaining their origins with XAI methods, practitioners can better understand its strengths and limitations and make informed decisions about its deployment. This process can also identify areas for improvement in the whole process, from the data collection to the model development and validation.

The PRE-ACT project will develop a fairness management pipeline, with the goal of detecting biases and providing explanations about their nature. The first fundamental component will be to stratify the patient data according to protected demographic attributes such as age, ethnicity, and income proxies - creating both basic and intersectional subgroups. The performance of the model will then be evaluated on all subgroups, so that disparities can be highlighted. This will allow to assess whether the model performs relatively poorly on subgroups of patients with specific demographic traits. Besides performance measure, the kind of per-subgroup ML misdiagnoses will be collected and analysed, thus providing a preliminary landscape of common misclassifications for all patient sub-cohorts.

Furthermore, explainability techniques will be exploited to obtain explanations about the aforementioned mistakes. As a result, both ML developer and clinical personnel will receive subgroup-level, human-understandable information about the driving causes of the model's erroneous outputs. This information can in turn be exploited to address the collected data, the training process, or the clinical reliability as a whole.

IV. Federated Learning

AI can be used with health data to improve clinical services and make health predictions. However, this kind of data

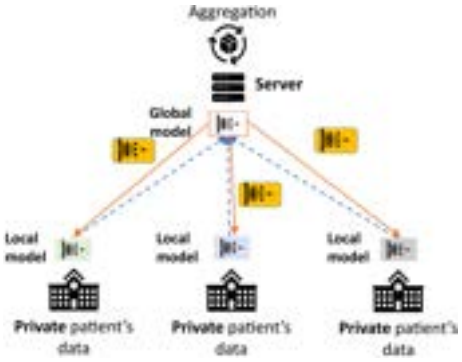

Fig. 1. The concept of FL. Participants share only model parameters and keep their sensitive data private.

is stored in various locations and cannot leave from these locations. Federated Learning (FL) is a distributed Machine Learning (ML) paradigm that enables the training of a global model on such private and decentralized data, by exchanging only models' parameters instead of the actual data samples [22]. FL consists of two phases (Fig. 1), which are repeated until convergence: (*i*) the *local training* phase, where each participant trains its model locally on its own data and updates the model's parameters by using, for example, stochastic gradient descent (SGD) and (*ii*) the *global aggregation* phase, where the updated parameters of the locally updated models from each participant are aggregated and averaged at a server, to obtain a new global model which in turn is sent back to all participants to initiate the next training round. In healthcare, FL is particularly useful as it allows for the creation of models that can be trained on patients' multi-modal data (e.g., medical images, genomics etc.) from multiple sources, such as hospitals and clinics, where regulatory policies or protection mechanisms for the data might exist [23].

Concretely, in the FL setting, hospitals can first leverage their local data to train their model and then share only their trained model' s weights with a server. There, these weights are aggregated (e.g., in the simplest form by averaging them) to create a global model, which in turn is sent back to all participants to initiate the next round. After consecutive rounds of updated weights exchange and aggregation, a global model is created that can for example predict which patients are at substantial risk of certain side effects, so that the most effective treatment can be decided. At the same time, the sensitive health-related data remain under the control of the individual hospitals, which helps to protect patients' privacy. Despite extensive research efforts, there are still two main challenges of FL in healthcare, in the context of hospital data, personalization and data heterogeneity and thus, the goal of PRE-ACT is to develop novel methods to address them.

A. Personalization

The plain FL scheme only develops a common output for all data holders (i.e., the hospitals) and therefore it does not adapt the model to each data holder's requirements. Specifically, in settings where the data are heterogeneous, e.g., non-independent and identically distributed (non-IID), the resulted global model obtained by minimizing the average loss (of all participants) could perform arbitrarily poorly once applied to the local dataset of each data holder. To this end, Personalized Federated Learning provides a solution as it allows for tailoring the machine learning model to the specific needs and characteristics of individual data holders. A state-of-the-art approach for addressing the challenges of personalization under non-iid data distributions is proposed in [24]. The authors separate the model into client-specific and shared parameters and perform an unbiased SGD step over both to find an exact solution to the optimization problem.

The dominant approach for personalization is local fine-tuning, where each client receives a global model and tunes it by using its own local data and performing several gradient descent steps. This approach is predominantly used in meta-learning methods such as model agnostic meta-learning [25] or Transfer Learning [26]. The main drawback of these methods is that the personalized model is bound to overfit as there is an inevitable tradeoff between personalization and global model performance. In addition, when data are heterogeneous among clients, using personalization in FL can be challenging and developing efficient methods for handling unbalanced data and class imbalance across clients is still an open problem. PRE-ACT will address these issues from various perspectives, for example an approach can be to train more than one model, such that each model is used to predict side effects for patients of a specific cohort.

B. Data heterogeneity in Federated Learning

Another challenge in FL is data heterogeneity which refers to the presence of dissimilar or diverse features in datasets or even to missing and inconsistent data among data holders. Especially in cases where data types differ, sharing a common model for each data holder is challenging. To address this issue, each client could have its own model parameters, rather than a shared model across all clients. Therefore, the model is able to adapt to the specific feature space of each data holder. For example, in [27] instead of sharing the model's parameters, they train locally a Generative Adversarial Network (GAN) and exchange synthetic images instead of model's weights. Unavoidably, such methods lead to inferior performance when compared against the standard FL training procedure, leaving a lot of room for improvement. PRE-ACT will cater to fill this gap, by developing methods that exploit the various types of non-iid hospital data under challenging privacy restrictions.

In realistic scenarios, clinical data are expected to have various forms like medical images, genomics' sequences etc. It is fairly common that for a number of patients some attributes might be missing from the datasets. This is a challenging aspect of FL, that despite research efforts, has not been addressed; leading to issues with convergence and accuracy. While data synthesis or augmentations techniques could be used, they provide only a superficial solution since the quality of the produced data is not on par with the actual data, especially in the medical domain. Participants (e.g., patients, hospitals etc.) with missing data may not be contributing equally to model updates, leading to bias and potential issues with fairness. These data heterogeneity issues in FL are aggravated in the case of medical data and the goal of PRE-ACT is to tackle them.

When data in hospitals differ not only in feature space but also in size, the naive parameters' aggregation would benefit

only the participants with more data locally, as it might result in a global model that is biased towards data holders with large datasets. Thus, it is of major importance that datasets are balanced in terms of size and features. Since, in medical imaging, the process of data acquisition and annotation is one of the most crucial and labor-intensive tasks, data synthesis and augmentation could be used as tools to mitigate this problem [28]. To this end, research should be directed towards aggregation techniques that take into consideration the difference between synthetic and actual data, when there are size differences in datasets. PRE-ACT will develop methods that incorporate the synthetic data in the training procedure of FL to cater for such cases.

V. Mobile and web applications

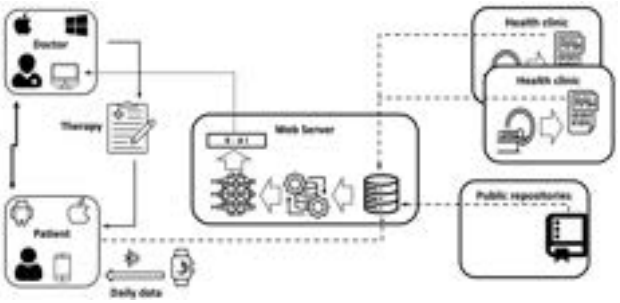

Fig. 2. System architecture.

The fourth technology of PRE-ACT is the design and development of mobile and web applications that emanate from a co-design approach with patients and physicians and aim at presenting the risk prediction to them. There exist several mobile app frameworks developed in the context of European projects that aid cancer patients such as FORTEe[2] and PREFERABLE.[3] These mostly motivate patients to improve quality of life. Another application, eSMART [29] proposes an intelligent symptom monitoring and alerting system. The user/patient fills symptom questionnaires, and she inserts her own health data, such as body temperature etc which are then transmitted to a central Web Server. The Web server employs simple NNs for each patient and classifies her symptoms. In case of unusual and dangerous symptoms, the system sends alerts to physicians and the hospital. ONCORELIEF [30] uses daily health data of patients recorded from smartwatches and informs them about their progress during the therapy period. The mobile device is used as data collector from smartwatch measurements and transmits them to a central data repository. ONCORELIEF deploys DNNs trained centrally with the newly received data and responds back with personalized suggestions or warnings to each patient, leveraging knowledge from similar patient behaviour.

The envisioned system architecture will consist of a Web Server (serving as the back-end), and two Front-end application versions; a Web/Desktop one for doctors and a Mobile one for patients. Details are shown in Fig. 2 and are discussed in detail in the sequel.

[2]https://fortee-project.eu/

[3]https://www.h2020preferable.eu/

A. System Architecture

1) Back-end: The Web Server (Back-end) includes big data storage (Database), processing, AI and communication (REST APIs) services. Each service is utilized by various software components. The central NoSQL database will store diverse medical data such as DICOM images and other data types. It will be updated with data from the three cohorts, e.g. auto-contouring data and daily data entered by each patient. A NoSQL database, such as MongoDB may store all data in a key-value fashion, providing quick data querying thanks to its efficient database indexing. The big data processing service, will include the Spark framework that is designed to execute SQL-like queries to NoSQL databases, and to efficiently share computation resources among computers to reduce execution time. Spark also includes ML libraries, such as MLib. For the AI service, relevant frameworks like TensorFlow, Keras, and PyTorch can be utilized. The DNNs can be trained continuously and achieve high accuracy since the database will be updated with fresh data from REST API calls of the Front-end applications. Finally, the communication service will include a powerful API, which consists of various endpoints and rules so as to exchange data with both front-end applications. REST API calls are executed from mobile devices to successfully retrieve data from the Web server over HTTP.

2) Front-end: Two versions of the same Front-end framework will be created, a Web application for doctors and a mobile app for patients. Both will include a storage component for storing health data and statistics into local database infrastructures such as SQLite and REALM. A content manager component includes various software utilities to display readable information to the user via functions of existing libraries, such as Charts library (Swift iOS). A synchronization component establishes secure connection with the Web Server by exchanging encrypted data through SSL/TLS protocols. An authentication component guarantees data privacy of each patient by integrating two-factor auth services to prevent unauthorized users. The mobile app may also include a health measurements component for extracting health data from third party services such as Health, Huawei Health etc that load health data from smartwatches. Users may also insert personalized data pertaining to lifestyle (e.g. smoking, exercise) and illnesses such as COVID-19, fever, allergies etc.

B. Mobile application challenges

One challenge associated with the mobile app is that of enhancing user familiarity, understanding and trust of the application. Both the risk prediction and its explanation will need to be adapted to user personal characteristics, such as age, education level, and specialization. The application may provide questionnaires to the user regarding the degree of understanding/explainability of each report and may need to adjust the explanation so as to increase user awareness. It is commonly understood that users tend trust more the applications that they understand [31]. Similar adaptations may be applied not only on the interfaces of how to communicate the risk and its explanation, but also to other interfaces of the app. Another issue is the user interface per se. The graphical user interface (GUI) should attract the user, be easy-to-use and create an engaging experience. Especially for elderly or impaired users, a friendly interface is necessary.

VI. Conclusion

We discussed four pillars underpinning *Trustworthy* AI and the way forward in the recently launched European Horizon Europe PRE-ACT project. Explainable AI and Fair AI will enable explainability of AI models and will uncover hidden biases in data. Federated Learning algorithms from decentralized data will train AI models while respecting data privacy. On the other hand, appropriately designed mobile applications will be the vehicles towards communicating the outcome of AI models to patients and physicians. Advances in these pillars will contribute to the grand objective of the PRE-ACT project, which is to communicate, in an explainable way, the risk prediction of radiotherapy side effects to patients and their clinicians so as to alleviate the occurrence of side effects and ultimately improve quality of life for cancer patients.

VII. Acknowledgment

This work was conducted in the context of the Horizon Europe project PRE-ACT (Prediction of Radiotherapy side effects using explainable AI for patient communication and treatment modification). It was supported by the European Commission through the Horizon Europe Program (Grant Agreement number 101057746), by the Swiss State Secretariat for Education, Research and Innovation (SERI) under contract number 22 00058, and by the UK government (Innovate UK application number 10061955).

References

[1] V. Belle and I. Papantonis, "Principles and practice of explainable machine learning," *Frontiers in Big Data*, vol. 4, 2021.

[2] M. T. Ribeiro, S. Singh, and C. Guestrin, ""why should I trust you?": Explaining the predictions of any classifier," in *Proceedings of the 22nd ACM SIGKDD Int.l Conf. on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016*, 2016, pp. 1135–1144.

[3] D. Slack, S. Hilgard, E. Jia, S. Singh, and H. Lakkaraju, "Fooling LIME and SHAP: Adversarial attacks on post hoc explanation methods," in *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, ser. AIES '20. New York, NY, USA: Association for Computing Machinery, 2020, p. 180–186.

[4] M. Golea, "On the complexity of rule extraction from neural networks and network querying," in *Rule Extraction From Trained Artificial Neural Networks Workshop, Society For the Study of Artificial Intelligence and Simulation of Behavior Workshop Series (AISB)*, 1996, pp. 51–59.

[5] R. Andrews, J. Diederich, and A. B. Tickle, "Survey and critique of techniques for extracting rules from trained artificial neural networks," *Knowledge-based systems*, vol. 8, no. 6, pp. 373–389, 1995.

[6] J. Diederich, *Rule extraction from support vector machines*. Springer Science & Business Media, 2008, vol. 80.

[7] L. Breiman, "Bagging predictors," *Machine learning*, vol. 24, no. 2, pp. 123–140, 1996.

[8] Y. Freund and R. E. Schapire, "A desicion-theoretic generalization of on-line learning and an application to boosting," in *European conference on computational learning theory*. Springer, 1995, pp. 23–37.

[9] G. Brown, J. Wyatt, R. Harris, and X. Yao, "Diversity creation methods: a survey and categorisation," *Information Fusion*, vol. 6, no. 1, pp. 5–20, 2005.

[10] G. Bologna, "Is it worth generating rules from neural network ensembles?" *Journal of Applied Logic*, vol. 2, no. 3, pp. 325–348, 2004.

[11] G. Bologna, "A rule extraction technique applied to ensembles of neural networks, random forests, and gradient-boosted trees," *Algorithms*, vol. 14, no. 12, p. 339, 2021.

[12] Z.-H. Zhou, Y. Jiang, and S.-F. Chen, "Extracting symbolic rules from trained neural network ensembles," *Artificial Intelligence Communications*, vol. 16, no. 1, pp. 3–16, 2003.

[13] U. Johansson, *Obtaining accurate and comprehensible data mining models: An evolutionary approach*. Linköping University, Department of Computer and Information Science, 2007.

[14] A. Hara and Y. Hayashi, "Ensemble neural network rule extraction using re-rx algorithm," in *Neural Networks (IJCNN), The 2012 International Joint Conference on*. IEEE, 2012, pp. 1–6.

[15] F. Jiang, Y. Jiang, H. Zhi, Y. Dong, H. Li, S. Ma, Y. Wang, Q. Dong, H. Shen, and Y. Wang, "Artificial intelligence in healthcare: past, present and future," *Stroke and vascular neurology*, vol. 2, no. 4, 2017.

[16] T. Davenoport and R. Kalakota, "The potential for artificial intelligence in healthcare," *Future healthcare journal*, vol. 6, no. 2, p. 94, 2019.

[17] D. Pedreshi, S. Ruggieri, and F. Turini, "Discrimination-aware data mining," in *Proc. of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2008, pp. 560–568.

[18] Z. Obermeyer, B. Powers, C. Vogeli, and S. Mullainathan, "Dissecting racial bias in an algorithm used to manage the health of populations," *Science*, vol. 366, no. 6464, pp. 447–453, 2019.

[19] D. K. Mulligan, J. A. Kroll, N. Kohli, and R. Y. Wong, "This thing called fairness: Disciplinary confusion realizing a value in technology," *Proc. ACM Hum.-Comput. Interact.*, vol. 3, 2019.

[20] C. Panigutti, A. Perotti, A. Panisson, P. Bajardi, and D. Pedreschi, "Fairlens: Auditing black-box clinical decision support systems," *Information Processing & Management*, vol. 58, no. 5, p. 102657, 2021.

[21] C. Panigutti, A. Beretta, F. Giannotti, and D. Pedreschi, "Understanding the impact of explanations on advice-taking: a user study for AI-based clinical decision support systems," in *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, 2022, pp. 1–9.

[22] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Artificial intelligence and statistics*. PMLR, 2017, pp. 1273–1282.

[23] M. J. Sheller, B. Edwards, G. A. Reina, J. Martin, S. Pati, A. Kotrotsou, M. Milchenko, W. Xu, D. Marcus, R. R. Colen *et al.*, "Federated learning in medicine: facilitating multi-institutional collaborations without sharing patient data," *Scientific reports*, vol. 10, no. 1, pp. 1–12, 2020.

[24] S. Nikoloutsopoulos, I. Koutsopoulos, and M. K. Titsias, "Personalized federated learning with exact stochastic gradient descent," *arXiv preprint arXiv:2202.09848*, 2022.

[25] Y. Jiang, J. Konečnỳ, K. Rush, and S. Kannan, "Improving federated learning personalization via model agnostic meta learning," *arXiv preprint arXiv:1909.12488*, 2019.

[26] Y. Chen, X. Qin, J. Wang, C. Yu, and W. Gao, "Fedhealth: A federated transfer learning framework for wearable healthcare," *IEEE Intelligent Systems*, vol. 35, no. 4, pp. 83–93, 2020.

[27] X. Cao, G. Sun, H. Yu, and M. Guizani, "Perfed-gan: Personalized federated learning via generative adversarial networks," *IEEE Internet of Things Journal*, 2022.

[28] M. Duan, D. Liu, X. Chen, R. Liu, Y. Tan, and L. Liang, "Self-balancing federated learning with global imbalanced data in mobile systems," *IEEE Transactions on Parallel and Distributed Systems*, vol. 32, no. 1, pp. 59–71, 2020.

[29] G. R.Maguire, L. McCann *et al.*, "Real time remote symptom monitoring during chemotherapy for cancer: European multicentre randomised controlled trial (esmart)," *Bmj*, vol. 374, 2021.

[30] J. Reis, L. Travado, T. Kosmidis *et al.*, "Oncorelief a digital guardian angel supported by an ai system to improve cancer patient quality of life, wellbeing and health outcomes: Protocol for a pilot study."

[31] J. Mirkovic, H. Bryhni, and C. M. Ruland, "Designing user friendly mobile application to assist cancer patients in illness management," in *The Third International Conference on eHealth, Telemedicine, and Social Medicine*, 2011, pp. 64–71.