Distributed learning for optimal radiomics knowledge

Please check the document version of this publication:

• A submitted manuscript is the version of the article upon submission and before peer-review. There can be important differences between the submitted version and the official published version of record. People interested in the research are advised to contact the author for the final version of the publication, or visit the DOI to the publisher's website.
• The final author version and the galley proof are versions of the publication after peer review.
• The final published version features the final layout of the paper including the volume, issue and page numbers.

Link to publication

Download date: 16 Aug. 2022

Distributed learning for optimal Radiomics knowledge

Fadila Zerka

Distributed learning for optimal Radiomics knowledge

DISSERTATION

to obtain the degree of Doctor at the Maastricht University,
on the authority of the Rector Magnificus, Prof. dr. Pamela Habibović
in accordance with the decision of the Board of Deans,
to be defended in public
on Wednesday 9 February 2022, at 13:00 hours

by
Fadila Zerka

Supervisor:
Prof. dr. Ph. Lambin

Co-supervisors:
Dr. H. C. Woodruff
Dr. S. Walsh (Radiomics, Liege, Belgium)

Assessment Committee:
Prof. dr. Manon van Engeland (Chair)
Dr. A. Lazovik (Institute for Mathematics and Computing Science, University
of Groningen)
Prof. dr. Benoît Macq (Catholic University of Louvain, Louvain-la-Neuve, Belgium)
Dr. M. Zamani Esteki

Contents

1

Introduction

1

Artificial intelligence (AI)

Artificial intelligence (AI) is a field of computer science, built to replicate or simulate the cognitive abilities of humans to learn. AI is powered by machine and deep learning algorithms running on powerful computers, intended to provide fast and accurate output [1].

Machine learning is a subfield of AI, that can detect and optimize for important features within data to characterize a predefined outcome. In the healthcare domain, data originate from a variety of sources:

- Clinical data: features derived from the routine clinical workflow (e.g., age, height, weight, etc.).
- Quantitative imaging data, otherwise called radiomics: that are derived from medical images, such as CT, MR, PET, and ultrasound. Radiomics features can be characterized into four feature types: intensity, shape, texture, and wavelet.
- Genomic data: derived from DNA/RNA information of an organism, mainly used to determine the functions of genes.

Machine learning can investigate each one of these data types individually or by coupling all or some of them [2,3]. In healthcare research, machine learning has greatly contributed to improving the diagnosis and prognosis processes by facilitating clinical decision-making [4,5].

Deep learning is a subfield of machine learning which functionality is inspired by the human brain. It uses neural networks with several embedded layers encapsulating a large quantity of neurons that can extract useful features from data. Different architectures have been proposed to perform different tasks and improve the performance [6]. Deep learning has gained great interest in healthcare research. In the field of radiology, the interest has mainly been to automate a multitude of manual processes such as image classification, segmentation, and detection and classification of abnormalities [7].

Distributed learning

Data is a key element to drive forward AI developments in both research and industry. In this context, the more quality data available to train and validate AI models, the more accurate is the outcome derived from it. For this reason, data collection and sharing are essential parts for every AI project. Nevertheless, recent legal and ethical considerations protecting personal data such as the European general data protection regulation (GDPR) and in USA the health insurance portability and accountability act (HIPAA) have made the data collection process challenging, time consuming, and costly [8,9]. Multicentric AI studies learning from distributed data has the potential to address these challenges. This

approach is known as distributed learning, where AI models characterized by machine learning or deep learning models learn iteratively or simultaneously from siloed data distributed across a network of partners [10–13]. Distributed learning enables large-scale data access while preserving privacy. The conventional distributed learning frameworks reported in the literature (prior to this thesis), all rely on a central server that coordinates the learning in the network. This design is subject to 1) a trust issue, as all partners are required to blindly trust the entity that manages the server, 2) a traceability issues that originates from the model centralization, thus the partners do not have a transparent monitoring system to evaluate the participation of the remaining network partners.

To overcome the issues associated with the conventional design of distributed learning, we propose a fully decentralized distributed learning framework. We achieve full decentralization by integrating a public blockchain (Ethereum), which is peer-to-peer (P2P) framework that enforces trust, immutability, transparency, traceability, and security on the network. In this thesis, we:

> Firstly, identify conventional distributed learning and Sequential learning coupled with blockchain technology namely Chained Distributed Machine learning (C-DistriM), as illustrated in Figure 1.1.
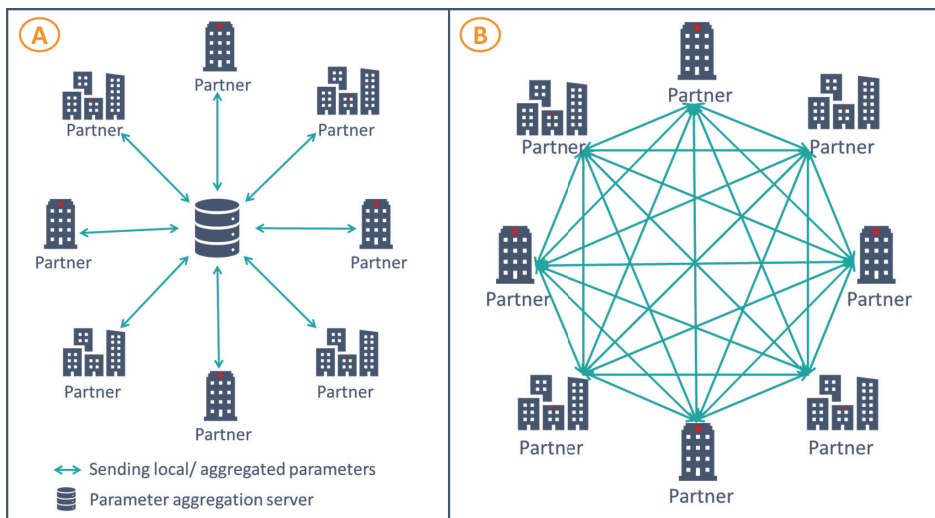
Figure 1.1: A) conventional distributed learning: all the partners are connected to a server that 1) initializes the learning, 2) each partner trains a portion of the model on local data, 3) each partner provides the model parameters to the server, 4) the server aggregates the parameters, 5) shares the updated parameters with the partners, 6) each partner retrains the local models using the updated parameters and sends them back to the server for update, 7) this process repeats until convergence criteria are met. **B)** decentralized sequential distributed learning: each partner holds a copy to the blockchain, that makes the learning history available for all, 1) training is initiated by the first partner that is connected to the system, 2) once the local model training is over, the next partner can start updating the previous model, 4) this process repeats until all partners finish updating the model.

Secondly, evaluate the application of sequential learning on small batch siloed data, then we evaluate the C-DistriM infrastructure in real world clinical settings.

The contribution of this work is to enable large-scale data access while preserving data privacy and provide a distributed learning framework that enforces trust, immutability, transparency, traceability, and security on the network.

Structure of the thesis

This thesis comprises the following topics:

1. **Centralized learning - radiomics:** presents a detailed overview of radiomics, in the standard centralized learning scheme, and analyzes its abilities in improving the clinical decision-making process (Chapter 2).
2. **Distributed learning:** appraises the existing implementations of conventional distributed learning frameworks and discuss possible future pathways (Chapter 3).
3. **Sequential learning and blockchain:** outlines blockchain technology and the important technical aspects of conventional distributed learning frameworks that must be adapted to achieve a successful integration with blockchain to accomplish full decentralization (Chapter 4).
4. **Evaluation:** is dedicated to first evaluate the fundamental change of the conventional distributed learning frameworks, represented by sequential distributed learning, on small batches of data (Chapter 5), second evaluate C-DistriM in a multinational setting (Chapter 6). Figure 1.2 shows a summary of the structure of this thesis.

This thesis is divided into six chapters.

Chapter 1 provides a general introduction to the thesis.

Chapter 2 describes the process of centralized radiomics, emphasizes its pitfalls, and challenges, opportunities within the context of improving clinical decision making.

Chapter 3 investigates the integration of distributed learning in healthcare, identify the limitations of the conventional distributed learning frameworks, and propose prospects to address those limitations.

Chapter 4 focuses on the implementation of a fully decentralized, immutable, and transparent distributed learning prototype (C-DistriM). The proposed distributed learning prototype is powered by the public Ethereum blockchain. The prototype was evaluated by simulating two distributed learning networks using two open-source datasets.

Chapter 5 is dedicated to investigating the potential of sequential distributed learning when applied on very small, siloed sets of clinical and imaging data. Using five open-source datasets, we evaluated three machine learning models: support vector machines, logistic regression, and perceptron. Additionally, we evaluated a more complex algorithm that consist of a convolutional neural network.

Chapter 6 presents the results of evaluating C-DistriM in a distributed learning network. The model consists of a convolutional neural network to estimate the cancer grade of prostate cancer patients using magnetic resonance imaging.

Chapter 7 discusses the work presented in this thesis and presents further perspectives.
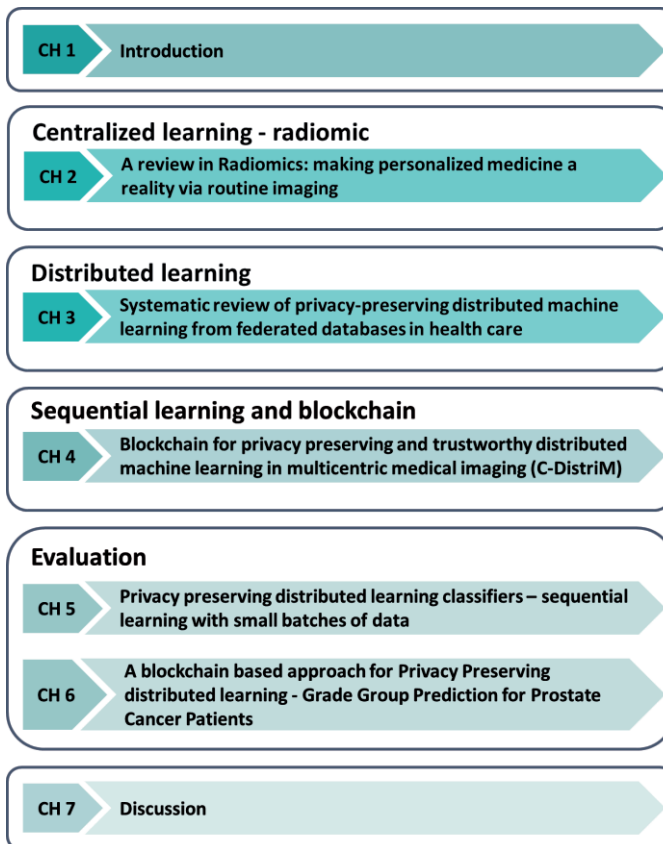

Figure 1.2: Summary of the thesis structure.

References

[1]   L. Zhang, Y. Pan, X. Wu, M.J. Skibniewski, Introduction to Artificial Intelligence, in: Artificial Intelligence in Construction Engineering and Management, Springer Singapore, Singapore, 2021: pp. 1–15. https://doi.org/10.1007/978-981-16-2842-9_1.

[2]   M. Kirienko, M. Sollini, M. Corbetta, E. Voulaz, N. Gozzi, M. Interlenghi, F. Gallivanone, I. Castiglioni, R. Asselta, S. Duga, G. Soldà, A. Chiti, Radiomics and gene expression profile to characterise the disease and predict outcome in patients with lung cancer, Eur J Nucl Med Mol Imaging. (2021). https://doi.org/10.1007/s00259-021-05371-7.

[3]   I. Shiri, M. Sorouri, P. Geramifar, M. Nazari, M. Abdollahi, Y. Salimi, B. Khosravi, D. Askari, L. Aghaghazvini, G. Hajianfar, A. Kasaeian, H. Abdollahi, H. Arabi, A. Rahmim, A.R. Radmard, H. Zaidi, Machine learning-based prognostic modeling using clinical data and quantitative radiomic features from chest CT images in COVID-19 patients, Computers in Biology and Medicine. 132 (2021) 104304. https://doi.org/10.1016/j.compbiomed.2021.104304.

[4]   Y. Zhou, X. Ma, T. Zhang, J. Wang, T. Zhang, R. Tian, Use of radiomics based on 18F-FDG PET/CT and machine learning methods to aid clinical decision-making in the classification of solitary pulmonary lesions: an innovative approach, Eur J Nucl Med Mol Imaging. (2021). https://doi.org/10.1007/s00259-021-05220-7.

[5]   P. Lambin, R.T.H. Leijenaar, T.M. Deist, J. Peerlings, E.E.C. de Jong, J. van Timmeren, S. Sanduleanu, R.T.H.M. Larue, A.J.G. Even, A. Jochems, Y. van Wijk, H. Woodruff, J. van Soest, T. Lustberg, E. Roelofs, W. van Elmpt, A. Dekker, F.M. Mottaghy, J.E. Wildberger, S. Walsh, Radiomics: the bridge between medical imaging and personalized medicine, Nat Rev Clin Oncol. 14 (2017) 749–762. https://doi.org/10.1038/nrclinonc.2017.141.

[6]   B. Tang, Z. Pan, K. Yin, A. Khateeb, Recent Advances of Deep Learning in Bioinformatics and Computational Biology, Front. Genet. 10 (2019) 214. https://doi.org/10.3389/fgene.2019.00214.

[7]   T. Wang, Y. Lei, Y. Fu, J.F. Wynne, W.J. Curran, T. Liu, X. Yang, A review on medical imaging synthesis using deep learning and its clinical applications, J Appl Clin Med Phys. 22 (2021) 11–36. https://doi.org/10.1002/acm2.13121.

[8]   M.A. Foy, General Data Protection Regulation May 2018 (GDPR): how does it affect us?, 7 (2018) 3.

[9]   C.G. Kels, HIPAA in the Era of Data Sharing, JAMA. 323 (2020) 476–477. https://doi.org/10.1001/jama.2019.19645.

[10]   M. Bogowicz, A. Jochems, T.M. Deist, S. Tanadini-Lang, S.H. Huang, B. Chan, J.N. Waldron, S. Bratman, B. O'Sullivan, O. Riesterer, G. Studer, J. Unkelbach, S. Barakat, R.H. Brakenhoff, I. Nauta, S.E. Gazzani, G. Calareso, K. Scheckenbach, F. Hoebers, F.W.R. Wesseling, S. Keek, S. Sanduleanu, R.T.H. Leijenaar, M.R. Vergeer, C.R. Leemans, C.H.J. Terhaard, M.W.M. van den Brekel, O. Hamming-Vrieze, M.A. van der Heijden, H.M. Elhalawani, C.D. Fuller, M. Guckenberger, P. Lambin, Privacy-preserving distributed learning of radiomics to predict overall survival and HPV status in head and neck cancer, Scientific Reports. 10 (2020) 4542. https://doi.org/10.1038/s41598-020-61297-4.

[11]   B.K. Beaulieu-Jones, W. Yuan, S.G. Finlayson, Z.S. Wu, Privacy-Preserving Distributed Deep Learning for Clinical Data, ArXiv:1812.01484 [Cs, Stat]. (2018). http://arxiv.org/abs/1812.01484 (accessed January 7, 2019).

[12]   P. Lambin, E. Roelofs, B. Reymen, E.R. Velazquez, J. Buijsen, C.M.L. Zegers, S. Carvalho, R.T.H. Leijenaar, G. Nalbantov, C. Oberije, M. Scott Marshall, F. Hoebers, E.G.C. Troost, R.G.P.M. van Stiphout, W. van Elmpt, T. van der Weijden, L. Boersma, V. Valentini, A. Dekker, 'Rapid Learning health care in oncology' – An approach towards decision support systems enabling customised radiotherapy', Radiotherapy and Oncology. 109 (2013) 159–164. https://doi.org/10.1016/j.radonc.2013.07.007.

[13]   M. Al-Shedivat, J. Gillenwater, E. Xing, A. Rostamizadeh, federated learning via posterior averaging: a new perspective and practical algorithms, (2021) 23.

Centralized learning - Radiomics

2

A review in Radiomics: making personalized medicine a reality via routine imaging

Julien Guiot | Akshayaa Vaidyanathan | Louis Deprez | Fadila Zerka |
Denis Danthine | Anne-Noelle Frix | Philippe Lambin | Fabio Bottari |
Nathan Tsoutzidis | Benjamin Miraglio | Sean Walsh | Wim Vos |
Roland Hustinx | Marta Ferreira | Pierre Lovinfosse |
Ralph T.H. Leijenaar

Abstract

Radiomics is the quantitative analysis of standard-of-care medical imaging; the information obtained can be applied within clinical decision support systems to create diagnostic, prognostic, and/or predictive models. Radiomics analysis can be performed by extracting hand-crafted radiomics features or *via* deep learning algorithms. Radiomics has evolved tremendously in the last decade, becoming a bridge between imaging and precision medicine. Radiomics exploits sophisticated image analysis tools coupled with statistical elaboration to extract the wealth of information hidden inside medical images, such as computed tomography (CT), magnetic resonance (MR) and/or positron emission tomography (PET) scans, routinely performed in the everyday clinical practice. Many efforts have been devoted in recent years to the standardization and validation of radiomics approaches, to demonstrate their usefulness and robustness beyond any reasonable doubts. However, the booming of publications and commercial applications of radiomics approaches warrant caution and proper understanding of all the factors involved to avoid "scientific pollution" and overly enthusiastic claims by researchers and clinicians alike. For these reasons the present review aims to be a guidebook of sort, describing the process of radiomics, its pitfalls, challenges and opportunities, along with its ability to improve clinical decision making, from oncology and respiratory medicine to pharmacological and genotyping studies.

Introduction

Imaging is a fundamental technology in medicine and is used in clinical practice to aid decision making for screening, diagnostic [1], therapeutic [2] and follow-up purposes. Radiomics was born in 2012 as an innovative approach to image analysis, using automated high-throughput extraction of large amounts of quantitative features from standard-of-care medical images [3,4]. The hypothesis is that quantitative analysis of medical image data can provide complementary information to aid physicians in the decision-making process, aided by automatic or semi-automatic software, in a fast and reproducible way [5]. Radiomics is the result of several decades of computer-aided diagnosis, prognosis, and therapeutics research [6,7]. A robust radiomics approach consists in the identification of a wide variety of quantitative features from medical images, the storage of such data in several independent databases functioning as a single entity (federated databases) [8] and the subsequent data mining to obtain clinically relevant outcomes [9]. Medical images such as CT, MR, and/or PET scans can be analysed and processed to extract relevant radiomics features which can be used for screening, diagnostic [10], follow-up and prognostic [11] purposes as well as for pharmacokinetic and pharmacodynamic studies [12–14]. Databases which collect and cross-reference vast amounts of radiomics data along with other relevant patient information from millions of cases are already a reality, but still present considerable management problems [15–18]. However, Radiomics is not the "philosopher stone" for clinical decision-making. Since its inception in 2012, the number of radiomics publications has grown exponentially (See Figure 2.1) as well as its detractors and disbelievers. The proven efficacy of radiomics approaches and the enthusiasm around this new method have to be tempered by its informed application and the careful evaluation of its real potential.
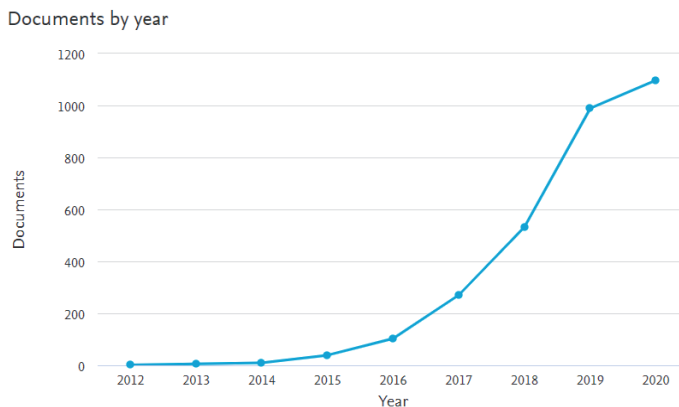

Figure 2.1: Number of "radiomics" publications per year (2012-2020). Data obtained from Scopus (09/09/2020).

Two main approaches are used for radiomics analysis, hand-crafted features and deep learning (DL). Radiomic hand-crafted features (such as intensity, shape, texture or wavelet) offer information on the specific area of the imaging scan one wishes to investigate, might be a tumour region or a whole organ. These features are distinct yet interconnected to other data sources (such as clinical, treatment or genomic data) [19]. The main challenge lies in the collection and integration of multimodal data sources in a quantitative fashion, delivering unambiguous clinical information and in turn allowing accurate and robust outcome prediction [20]. Deep learning methods instead uses a data driven approach for model creation, mimicking simplified brain neuron interactions. Deep learning has the advantage of not needing prior segmentation of the imaging scan: however the "black box" approach of DL i.e. the lack of interpretability of the models and features generated is seen as the main limitations for clinical applicability. Moreover, DL approaches need a large amount of data to truly express their potential, and sometimes the patients cohort available, for example in case of rare diseases, are not enough to leverage a DL architecture in an effective manner.

For as much as this scenario seems straightforward and most alluring for clinicians, there are still too many published prediction models which lack standardized evaluation of their performance, reproducibility, and/or clinical utility [21,22].

In this review, the pitfalls and challenges along with the opportunities presented by radiomics to improve personalized precision medicine will be showcased, stressing important methodological aspects of radiomics prediction model both in term of development and validation. We will explore the advanced information technologies that are essential for the simultaneous management of radiomics and clinical data. Finally, we will present our outlook on the necessary steps that still need to be taken to ensure widespread acceptance of radiomics in current clinical practice.

Good practices in radiomics studies

Radiomics can be defined as a collection of methods (algorithms) used to extract a large number of features from radiographic medical images [9]. Radiomics emerged originally in the field of oncology [1,23]; however, it can be applied to any medical study where a disease or a condition can be imaged [24–27]. A radiomics study can be divided into four main phases: data selection and curation, features extraction, exploratory analysis and modelling. Below we report a typical step-by-step radiomics workflow (Figure 2.2).
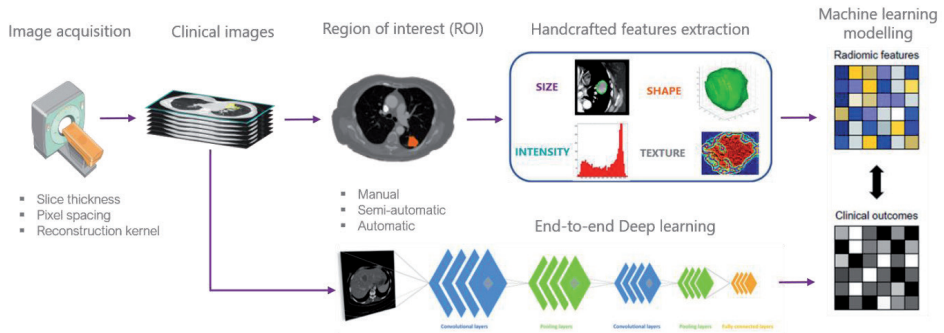
Figure 2.2: Scheme of the radiomics workflow for hand-crafted features (top) and deep learning (bottom).

Data selection and curation

Radiomic analyses starting point is the selection of an imaging technique (CT, MRI, PET, etc.), the identification of the region or volume of interest (ROI or VOI) and the choice of a specific prediction target — the relevant clinical question that the radiomics analysis aims to answer. In a typical oncological study, the whole primary tumour is analysed and linked to available data on treatment outcomes and disease prognosis, such as survival rate or tumour shrinkage. Radiomic analyses can be performed on subregions of the tumour (habitats), metastatic lesions, as well as in normal tissues. Radiomics analysis, however, is not restricted to radiotherapy and can be applied to any image generated in the clinical setting [28–30]. The use of standardized imaging protocols to eliminate unnecessary variability is of paramount importance [9,31] and has been recognized through the years as one of the main factors leading to low quality radiomics analysis [32]. Still nowadays, however, non-standardized imaging protocols are commonplace: reproducibility and comparability of radiomic studies would immensely benefit from clear guidelines on how imaging protocols should be applied and reported. To at least partially overcome these issues, images datasets must be carefully evaluated and where possible standardized, following well-established radiomics criteria [33]. Selection of slice increment and reconstruction kernel can be used, for example, as criteria to include or exclude imaging scans for radiomics feature extraction.

- **Medical imaging**

Segmentation

Segmentation is the first fundamental step in radiomics analysis and can be performed manually by expert radiologist/clinicians or (semi-) automatically [34]. Both approaches have their pros and cons and the most suited one varies on a case-by-case basis [35,36]. In general, automatic segmentation is more reproducible and faster than hand-made segmentation. The segmentation step determines which voxels within an image are analysed: it is easy to see that the variability in segmentation (both human- and

machine-driven) can introduce bias in the evaluation of the derived radiomic features [37]. For example, a semi-automatic segmentation method can result in different radiomic features than a manual segmentation, as well as segmentation performed by two different physicians. Comparing multiple segmentation approaches might be a solution to limit the amount of this bias [38] . It may consist in different clinical experts manual segmentation or the, perturbation of the automatic or semi-automatic segmentations with noise [39] or again the combination of different segmentation algorithms [40,41].

Phantom studies and feature stability

Another source of variability in the preliminary radiomics phase is the inter-machine and inter-vendor differences between the scanners employed [42]. In most real-life situation, the radiomics study must rely on data acquired on different scanners from different producers and with different "history": thus, not taking into account this systematic source of uncertainty might jeopardize the radiomics model prediction capabilities. To overcome at least part of this intrinsic limitation, the use of phantoms (i.e. an object built in shape and materials as close as possible to human tissue and organs) is a suitable mean to assess and account for the possible similarities and difference [43]. Radiomics features need also to be robust with respect to other possible sources of variability such as target volume motion, expansion or shrinkage. To probe the feature resilience, test-retest approaches [44,45] can be exploited to measure feature stability: for example, two datasets of images acquired at two different time points from the same subjects (e.g. patients or phantoms) or the use of cohorts from multiple sources [46,47]. In this way, volatile or robust features can be identified and excluded from model development.

To ameliorate the reproducibility of radiomics features, several methods of harmonization have been proposed in the literature. The ComBat method, initially developed for genomics, aims to remove non-biological differences related to scanner type in an effort to combine radiomics features extracted from data coming from different centres [32,48,49]. Other methods includes Neural Network training for radiomics feature standardization [50], intensity and diffusion maps harmonization [47,51] and data augmentation with generative adversarial networks (GAN) [52]. For a complete overview see [53].

Feature extraction

The essence of radiomics is the extraction of quantitative image features to characterize VOIs. Hand-crafted radiomics feature can be divided in five group: size and shape based–features, descriptors of the image intensity histogram, descriptors of the (spatial) relationships between image voxels, features extracted from filtered images, and fractal features [54,55]. Feature values are dependent on image pre-processing

steps performed, such as filtering, or intensity discretization and reconstruction. Furthermore, variations exist in feature nomenclature, mathematical definition, extraction methodology, and software implementation of the extraction algorithms [56–58]. To harmonize radiomics features and model reports, all these differences has to be taken into account and clear specification included with each model [33].

Exploratory analysis

The true potential of radiomics approaches lies in the possibility to combine radiomics and other non-image based features with the prediction endpoint to create a single dataset. This approach allows the evaluation of possible correlations between features. However, some radiomics features that are highly correlated with other routine clinical features (such as tumour stage or age) might not provide additional meaningful information. Approaches such as (unsupervised) clustering, PCA (principal component analysis) [59] or MRMR (maximum relevance minimum redundancy) [60] identify and eliminate redundancy, for instance, by reducing highly correlated features to a single representative archetypical feature. This is a fundamental step to avoid overfitting [61,62]. On the other hand, additional data collected, for example, from multiple segmentations or phantom studies can be used to test the feature robustness [63,64]. This process of reduction should be described clearly, to avoid misinterpretation and aid in the unambiguous identification of relevant features. Also, univariable correlations of single radiomics features with clinical outcome is part of the exploratory analysis and could inform the subsequent modelling step, underlining relations between single radiomics features with clinical covariates of interest.

Modelling

After features extraction and possible reduction, the creation of the radiomic model encompasses three major steps: feature selection, modelling methodology, and validation. Regarding the choice of modelling methodology, the identification of the best machine-learning method is a crucial step; thus, in an ideal scenario, multiple methods should be utilized and compared [65] and their implementation should be comprehensively documented. Another fundamental point in the modelling phase is the validation, which has to be performed in order to verify the applicability of the model in a real-world situation. Ideally, the model should be internally and externally validated, and the performance compared and clearly reported [20,66,67].

- **Feature selection**

The number of radiomics feature which could be extracted from medical images is technically unlimited. Several different filters, feature categories and other parameters can be used to mine the information hidden inside an imaging scan. Including all the possible features, even if practically possible, would result in overfitting which in turn renders the model useless for patients not previously evaluated (the so called curse

of dimensionality) [68,69]. The most used approach is the reduction to archetypal feature representing a group or class of features, identified by dimensionality reduction techniques. Several different kind of clustering algorithms and PCA are available and also this choice has to be justified and reported in detail, to promote transparency and replicability. Again, the same feature might be relevant for a given dataset, segmented in a certain way for a specific end-point prediction but not important whit a different segmentation routine or a different cohort of patients.

- **Modelling methodology**

The choice of modelling technique has been proven to affect prediction performance in radiomics [65]. Ideally, multiple modelling methodologies should be tested to select the best approach for the given data set and the other parameters involved in the creation of the model. Comparison between machine Learning (ML) and deep Learning (DL) approaches are common [70,71] or even combination of both [72] and the final choice has to take into account, along with the performance of the model, also the applicability of the proposed strategy in a real-world situation, considering for example computational burden or explainability of the resulting predictions [73,74]. Another key point in the selection of modelling methodology is replicability by other researchers, in the light of responsible and transparent research and innovation. This can be achieved, for example, by making the software code available in public repositories such as GitHub [75], Gitlab [76] and OpenML [77]. Also, many scientific journals put in place, in the last years, tools to help data and algorithms sharing, making these available to the scientific community.

- **Validation**

Validation techniques are needed to assess the generalizability of the model predictions. Validation answers the question whether the model is predictive for the whole target patient population or just for a particular subset of cases analysed. Model performances are typically measured in terms of *discrimination* and *calibration*. Discrimination is represented by the concordance statistics. For example the discrimination metric for binary outcome is the receiver operating characteristic (ROC) curve, or area under the ROC curve (AUC) [78]. The AUC is linked to the sensitivity and specificity of the model and represents the probability that a random patient matching an outcome is assigned in the class specific for that outcome with a larger probability than another random patient who does not match the outcome. The calibration, instead, is a measure of the agreement between observed outcomes and model predictions [79]. Calibration can be reported using a calibration plot and calibration-in-the-large/slope, with the Brier score, the mean squared prediction error, as a measure of overall performances.

The statistical methods used on both training and validation data sets need to be reported in detail. A valid model must exhibit statistical consistency between the training and validation sets. In terms of validation set selection, an externally validated

model has more credibility than an internally validated one because validation with independent data sets is considered more robust [67,79–81]. For "good radiomics practice", the reproducibility and replicability of the model should also be included in the validation step. Reproducibility relates to the verification of the result by independent researchers using the same methodology and data set, to verify the absence of errors, while replicability means the possibility of replicating the radiomics analysis with the same methodology but different appropriate datasets, to generalize the original findings [82–86]. Reproducibility and replicability in radiomics are, however, not possible if researchers do not disclose all the details of the analysis performed. Each radiomics model must be accompanied by the of imaging protocol used for image collection, selected scans for analysis with exclusion and inclusion criteria, segmentations of VOIs, detailed accounts of how features were extracted (including the pre-processing and feature reduction , and of the modelling methodology used (ideally, the code) [87].

The Radiomics Quality Score: 3 years later

In 2017 the Radiomics Quality Score (RQS) was proposed in an effort to help the scientific community assessing the quality and scientific/clinical value of a radiomics study at a glance [4]. A similar example is the Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD) initiative [80]. The RQS is determined by 16 key criteria which are assigned to a point value for a maximum of 36 points (100%). These criteria cover image acquisition protocols, statistic data treatment, cohorts provenance and open science policies, encompassing all the relevant aspects which a reliable radiomics publication should present. Evaluating, 3 years after its publication, on what has been achieved with this initiative, we can clearly see that the road ahead is still long. The RQS was espoused with enthusiasm by the scientific community: however, was mainly used to assess the quality of already published studies. Recently, several systematic reviews have appeared, covering a variety of cancers: breast cancer [88,89], hepatocellular carcinoma [90], gliomas [91], prostate cancer [92], lung cancer [93,94] and renal carcinoma [95]. All authors reported a very low RQS for the investigated publications (lower than 50%, and in most cases lower than 20%) with lack of external validation, prospective study registration and feature robustness test as the main causes. Also, open science policies and relevant clinical outcomes are among the most critical points raised for already published radiomics studies. To the best of our knowledge, however, researchers are still reticent in calculating and publishing the RQS for their newly developed radiomics models. In our opinion this is slowing down considerably the acceptance of radiomics as a full-fledged ancillary method for critical clinical decision-making. The RQS should be seen as a quality seal of the published results more than a way of underlining the possible weaknesses of the proposed model. A low RQS score does not necessarily mean that the research is not sound but it should be part of the

discussion and motivate the authors to carefully think through all the steps performed and discuss why deviation from "good practice" is justified or what can be done better. Editors, reviewers, and readers should be able to ascertain whether a radiomic study is compliant with good radiomics practice or, alternatively, whether the authors have justified any non-compliance. Overly optimistic claims concerning robustness and generalizability diminish scientific and clinical impact and should be avoided.

The consideration about the apparently limited impact of the RQS thus far might spur the scientific community to ask the question whether the RQS requires regular evaluation and updating in such a fast-evolving field, like radiomics. For example, RQS was tailored on hand-crafted feature while nowadays deep learning is gaining momentum and some of its inherent strengths and weaknesses might not be captured correctly in the current version of the RQS score.

Radiomics towards personalized medicine

Virtual biopsy
In patients with cancer, different parts of the tumour have distinct molecular characteristics, but also different lesions (metastases) from a tumour disease, which may have a role in terms of therapeutic efficacy and such differences, might change over time. As it is not possible to take samples of every part of each tumour at multiple time points, the optimal characterization of tumours is not achieved using biopsy[96]. However, radiomics might be used to "sample" different part of the tumour at different time points (i.e. different scans) and, along with genomic data, used as a virtual biopsy tool [97,98]. The combination of radiomics and genomics is called radiogenomics and has gathered considerable attention in the last years, as a way of augmenting the power of both approaches, for personalized medicine and treatment follow-up [99–101].

Beyond oncology
Radiomics was mostly employed in oncology up to now, but in the last years showed its potential for other clinical applications. Radiomics analysis was performed on MRI scans to distinguish between different cognitive disorders such as Alzheimer's disease [102–104], autism spectrum disorder [105] or amnestic mild cognitive impairment (aMCI) [106,107]. Another field in which radiomics might give a relevant contribution is bone disease study. Radiomics methods have been reported for the early identification of osteoporosis [108] or for classification of osteoporotic patients compared to normal subjects or suffering from osteopenia [109]. Radiomics was also applied to maxillofacial radiology [110]. Among the new application of radiomics, the 2020 brought forth a clear winner. The outbreak of new SARS-CoV-2 virus and the subsequent pandemic placed the research community under unprecedented pressure in the race to find better diagnostic

and therapeutic tools to fight this threat to human health. The correlation between COVID-19 infection and lung CT scans characteristics was reported early during the year [111,112], suggesting that radiomics might be a successful approach in the early diagnosis and prognosis of COVID-19 patients. In the last months of 2020, the scientific community has produced a deluge of radiomics publications related to COVID-19. Radiomics has been used for diagnosis [41,113,114], to distinguishing pulmonary infections from different sources [115–117] or predicting the length of hospitalization [118]. A complete overview of all the radiomics approaches for COVID-19 is beyond the scope of the present review; a relevant collection of the most interesting articles can be found in [119,120].

Delta radiomics

The vast majority of radiomics methods published focus on imaging data acquired at a single time point, mostly imaging tumours before the start of treatment. Delta-radiomics introduces a time component with the extraction of quantitative features from image sets acquired over the course of treatment [121–123], which provides information on the evolution of feature values. Delta-radiomics promises to improve diagnosis, prognosis, prediction, monitoring, image-based intervention, or assessment of therapeutic response [124,125]. Delta radiomics has been proven effective in the study of immunotherapy response [126,127] or to predict recurrence in oncological patients [128].

Open science and data sharing

There is a pressing need to embrace knowledge and data-sharing technology [129], which transcends institutional and national boundaries [130]. This is especially true for radiomics whose potency is directly linked to the amount and quality of data available. Larger datasets, deeper clinical and molecular information and homogeneous imaging sources will result in more robust and reliable radiomics models. To unlock the full potential of radiomics for clinical decision making, the research and clinical community must strive for a truly open science – sharing datasets, algorithms, best practices and finding new ways to improve collaborations. One initiative to accomplish these goals is CancerLinQ [131], the ASCO data centralization approach. Another initiatives are worldCAT and its European counterpart euroCAT [132] that consist in a novel data-federated approach that successfully links radiotherapy institutes in the Netherlands, Germany, Belgium, Italy, Denmark, Australia, China, India, South Africa, Ireland, UK, USA and Canada [133,134]. Other important initiatives include The Cancer Imaging Archive (TCIA) [135], The quantitative imaging network (QIN) [136], the quantitative imaging biomarkers alliance (QIBA) [137], the MEDomics consortium [138], and quantitative imaging in cancer: connecting cellular processes with therapy (QuIC-ConCePT) [139]. To overcome data sharing issues, such as privacy concerns or insufficient infra-

structures, an approach based on distributed machine learning for radiomics model's creation has been proposed. The concept of distributed (federated, privacy preserving) machine learning is not new in healthcare application [8,133] but has recently shown its potential for radiomics [140,141]. For example, Shi et al. performed a multi-centre study to develop a radiomic signature for lung cancer in one institution and validated the performance in an independent institution, without the need for data exchange [142]. In another recent case study, Bogowicz et al. developed and validated a radiomic signature for head and neck cancer, training the model remotely from 6 independent cohorts, showing that the performances of the distributed model were as good as the one obtained with traditional radiomic approach [143]. The next step in this open science initiative for radiomics should be the creation of database to store and cross-reference radiomics feature and relevant clinical data (radiomics ontology [144,145]). Also, the accessibility of radiomics in general must be improved and some initiatives in this regard are already in place, especially from a software perspective. Several open source or freeware software are already available [58,146–148] and code sharing is becoming more and more accepted in the scientific community.

New applications of radiomics

In the last years, radiomics had broaden its horizon, pushing the boundaries of what was achieved thus far. The possibility to use radiomics signatures to explore new medical conditions is expanding. Recently, a signature for the determination of chromosome deletion in low grade glioma (LGG) patients [149] has been reported. The authors developed two radiomics signature, composed of seven and five features respectively, extracted from T2- and T1-weighted post-contrast MRI. Both signatures showed an accuracy higher than 0.70. Another recent example of the novel pathways for radiomics is represent by the paper of Mu et al. [150]. The authors realized a radiomics signature on baseline PET-CT image of NSCLC patients treated with immune checkpoint inhibitors. The signature was able to predict prolonged weight loss syndrome (cachexia), which contribute to primary resistance to immune checkpoint inhibitors therapy. The signature was also able to predict durable clinical benefit (DCB), progression-free survival (PFS) and overall survival (OS). The performance on this signature in the external testing cohort were satisfying (AUC higher than 0.65). Radiomics was also used to explore different imaging technique such as ultrasound imaging (US). Chiappa et al. [151] developed a signature for differential diagnosis of myometrial tumors. While the study is based on single center data, performance of the signature in discriminating between patients with sarcoma or myoma were very good (accuracy of 0.85), proving that also US imaging can be successfully mined *via* radiomics approaches. The combination of radiomics and digital pathology has been also recently reported for different kind of cancer such as nasopharyngeal carcinoma [152], non-small cell lung cancer [153] and pancreatic cancer [154] among others. All these examples showcase the versatility

of radiomics which could stem from a closer collaboration with clinical expert and AI scientist beyond the current applications.

Conclusions and future outlook

The outlook for radiomics is very promising and the efforts devoted to its standardization are already bearing their fruits. However, there is still much work to do, especially to link fundamental research to current clinical practice. Physicians and healthcare personnel should be involved from the start of the process, along with relevant authorities. On the other hand, more effort should be devoted in the technological transfer, taking the published research and perform the necessary steps to bring it from a (validated) proof-of-concept to the clinic. This also emphasizes the need for comprehensive and universal indicators (such as the RQS) of the quality of a model. The normative framework is currently evolving along with the innovations in the field of AI driven healthcare. For example, FDA is gathering feedbacks and propositions to draft a novel regulatory framework for AI/ML based medical devices [155]. Paradigms need to re-invented to allow these breakthroughs to reach the clinic in the very near future, always putting patient's welfare first. Personalized, patient-centric medicine is almost a reality and radiomics is playing a major role in it and will represent one of the key factors for the future of healthcare.

References

[1]   H.J.W.L. Aerts, E.R. Velazquez, R.T.H. Leijenaar, C. Parmar, P. Grossmann, S. Carvalho, J. Bussink, R. Monshouwer, B. Haibe-Kains, D. Rietveld, F. Hoebers, M.M. Rietbergen, C.R. Leemans, A. Dekker, J. Quackenbush, R.J. Gillies, P. Lambin, Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach., Nat. Commun. 5 (2014) 4006. https://doi.org/10.1038/ncomms5006.

[2]   L. Hood, S.H. Friend, Predictive, personalized, preventive, participatory (P4) cancer medicine, Nat. Rev. Clin. Oncol. 8 (2011) 184–187. https://doi.org/10.1038/nrclinonc.2010.227.

[3]   P. Lambin, E. Rios-Velazquez, R. Leijenaar, S. Carvalho, R.G.P.M. Van Stiphout, P. Granton, C.M.L. Zegers, R. Gillies, R. Boellard, A. Dekker, H.J.W.L. Aerts, Radiomics: Extracting more information from medical images using advanced feature analysis, Eur. J. Cancer. 48 (2012) 441–446. https://doi.org/10.1016/j.ejca.2011.11.036.

[4]   P. Lambin, R.T.H. Leijenaar, T.M. Deist, J. Peerlings, E.E.C. De Jong, J. Van Timmeren, S. Sanduleanu, R.T.H.M. Larue, A.J.G. Even, A. Jochems, Y. Van Wijk, H. Woodruff, J. Van Soest, T. Lustberg, E. Roelofs, W. Van Elmpt, A. Dekker, F.M. Mottaghy, J.E. Wildberger, S. Walsh, Radiomics: The bridge between medical imaging and personalized medicine, Nat. Rev. Clin. Oncol. 14 (2017) 749–762. https://doi.org/10.1038/nrclinonc.2017.141.

[5]   V. Kumar, Y. Gu, S. Basu, A. Berglund, S.A. Eschrich, M.B. Schabath, K. Forster, H.J.W.L. Aerts, A. Dekker, D. Fenstermacher, D.B. Goldgof, L.O. Hall, P. Lambin, Y. Balagurunathan, R.A. Gatenby, R.J. Gillies, Radiomics: the process and the challenges, Magn. Reson. Imaging. 30 (2012) 1234–1248. https://doi.org/https://doi.org/10.1016/j.mri.2012.06.010.

[6]   AA.VV., Medicine: Computers by the Bedside, Nature. 224 (1969) 636–637. https://doi.org/10.1038/224636b0.

[7]   H.M. Schoolman, L.M. Bernstein, Computer use in diagnosis, prognosis, and therapy, Science (80-.). 200 (1978) 926 LP – 931. https://doi.org/10.1126/science.347580.

[8]   F. Zerka, S. Barakat, S. Walsh, M. Bogowicz, R.T.H. Leijenaar, A. Jochems, B. Miraglio, D. Townend, P. Lambin, Systematic Review of Privacy-Preserving Distributed Machine Learning From Federated Databases in Health Care, JCO Clin. Cancer Informatics. (2020) 184–200. https://doi.org/10.1200/cci.19.00047.

[9]   R.J. Gillies, P.E. Kinahan, H. Hricak, Radiomics: Images Are More than Pictures, They Are Data, Radiology. 278 (2015) 563–577. https://doi.org/10.1148/radiol.2015151169.

[10]   Z. Liu, S. Wang, D. Dong, J. Wei, C. Fang, X. Zhou, K. Sun, L. Li, B. Li, M. Wang, J. Tian, The applications of radiomics in precision diagnosis and treatment of oncology: Opportunities and challenges, Theranostics. 9 (2019) 1303–1322. https://doi.org/10.7150/thno.30309.

[11]   B.-H. Zheng, L.-Z. Liu, Z.-Z. Zhang, J.-Y. Shi, L.-Q. Dong, L.-Y. Tian, Z. Ding, Y. Ji, S.-X. Rao, J. Zhou, J. Fan, X.-Y. Wang, Q. Gao, Radiomics score: a potential prognostic imaging feature for postoperative survival of solitary HCC patients, BMC Cancer. 18 (2018) 1148. https://doi.org/10.1186/s12885-018-5024-z.

[12]   S. Monti, M. Aiello, M. Incoronato, A.M. Grimaldi, M. Moscarino, P. Mirabelli, U. Ferbo, C. Cavaliere, M. Salvatore, DCE-MRI pharmacokinetic-based phenotyping of invasive ductal carcinoma: A radiomic study for prediction of histological outcomes, Contrast Media Mol. Imaging. 2018 (2018). https://doi.org/10.1155/2018/5076269.

[13]   J.E. Bibault, P. Giraud, C. Durdux, J. Taieb, A. Berger, R. Coriat, S. Chaussade, B. Dousset, B. Nordlinger, A. Burgun, Deep Learning and Radiomics predict complete response after neo-adjuvant chemoradiation for locally advanced rectal cancer, Sci. Rep. 8 (2018) 1–8. https://doi.org/10.1038/s41598-018-30657-6.

[14]   X. Ii Song, J.L. Ren, D. Zhao, L. Wang, H. Ren, J. Niu, Radiomics derived from dynamic contrast-enhanced MRI pharmacokinetic protocol features: the value of precision diagnosis ovarian neoplasms, Eur. Radiol. I (2020). https://doi.org/10.1007/s00330-020-07112-0.

[15]   E. Roelofs, A. Dekker, E. Meldolesi, R.G.P.M. van Stiphout, V. Valentini, P. Lambin, International data-sharing for radiotherapy research: An open-source based infrastructure for multicentric clinical data mining, Radiother. Oncol. 110 (2014) 370–374. https://doi.org/10.1016/j.radonc.2013.11.001.

[16]   E. Roelofs, L. Persoon, S. Nijsten, W. Wiessler, A. Dekker, P. Lambin, Benefits of a clinical data warehouse with data mining tools to collect data for a radiotherapy trial, Radiother. Oncol. 108 (2013) 174–179. https://doi.org/10.1016/j.radonc.2012.09.019.

[17]   R. Miotto, L. Li, B.A. Kidd, J.T. Dudley, Deep Patient: An Unsupervised Representation to Predict the Future of Patients from the Electronic Health Records, Sci. Rep. 6 (2016) 26094. https://doi.org/10.1038/srep26094.

[18]   K.T. Nead, G. Gaskin, C. Chester, S. Swisher-McClure, J.T. Dudley, N.J. Leeper, N.H. Shah, Androgen Deprivation Therapy and Future Alzheimer's Disease Risk, J. Clin. Oncol. 34 (2015) 566–571. https://doi.org/10.1200/JCO.2015.63.6266.

[19]   R.A. Gatenby, O. Grove, R.J. Gillies, Quantitative Imaging in Cancer Evolution and Ecology, Radiology. 269 (2013) 8–14. https://doi.org/10.1148/radiol.13122697.

2

[20] P. Lambin, J. Zindler, B.G.L. Vanneste, L. Van De Voorde, D. Eekers, I. Compter, K.M. Panth, J. Peerlings, R.T.H.M. Larue, T.M. Deist, A. Jochems, T. Lustberg, J. van Soest, E.E.C. de Jong, A.J.G. Even, B. Reymen, N. Rekers, M. van Gisbergen, E. Roelofs, S. Carvalho, R.T.H. Leijenaar, C.M.L. Zegers, M. Jacobs, J. van Timmeren, P. Brouwers, J.A. Lal, L. Dubois, A. Yaromina, E.J. Van Limbergen, M. Berbee, W. van Elmpt, C. Oberije, B. Ramaekers, A. Dekker, L.J. Boersma, F. Hoebers, K.M. Smits, A.J. Berlanga, S. Walsh, Decision support systems for personalized and participative radiation oncology, Adv. Drug Deliv. Rev. 109 (2017) 131–153. https://doi.org/https://doi.org/10.1016/j.addr.2016.01.006.

[21] A.J. Vickers, Prediction Models: Revolutionary in Principle, But Do They Do More Good Than Harm?, J. Clin. Oncol. 29 (2011) 2951–2952. https://doi.org/10.1200/JCO.2011.36.1329.

[22] E.J. Limkin, R. Sun, L. Dercle, E.I. Zacharaki, C. Robert, S. Reuzé, A. Schernberg, N. Paragios, E. Deutsch, C. Ferté, Promises and challenges for the implementation of computational medical imaging (radiomics) in oncology, Ann. Oncol. 28 (2017) 1191–1206. https://doi.org/10.1093/annonc/mdx034.

[23] R.T.H. Leijenaar, M. Bogowicz, A. Jochems, F.J.P. Hoebers, F.W.R. Wesseling, S.H. Huang, B. Chan, J.N. Waldron, B. O'Sullivan, D. Rietveld, C.R. Leemans, R.H. Brakenhoff, O. Riesterer, S. Tanadini-Lang, M. Guckenberger, K. Ikenberg, P. Lambin, Development and validation of a radiomic signature to predict HPV (p16) status from standard CT imaging: A multicenter study, Br. J. Radiol. 91 (2018) 20170498. https://doi.org/10.1259/bjr.20170498.

[24] A.F. Leite, K. de F. Vasconcelos, H. Willems, R. Jacobs, Radiomics and Machine Learning in Oral Healthcare, Proteomics - Clin. Appl. 14 (2020). https://doi.org/10.1002/prca.201900040.

[25] H. Sun, Y. Chen, Q. Huang, S. Lui, X. Huang, Y. Shi, X. Xu, J.A. Sweeney, Q. Gong, Psychoradiologic utility of MR imaging for diagnosis of attention deficit hyperactivity disorder: A radiomics analysis, Radiology. 287 (2018) 620–630. https://doi.org/10.1148/radiol.2017170226.

[26] P. Lovinfosse, D. Visvikis, R. Hustinx, M. Hatt, FDG PET radiomics: a review of the methodological aspects, Clin. Transl. Imaging. 6 (2018) 379–391. https://doi.org/10.1007/s40336-018-0292-9.

[27] L. Sibille, R. Seifert, N. Avramovic, T. Vehren, B. Spottiswoode, S. Zuehlsdorff, M. Schäfers, 18F-FDG PET/CT Uptake Classification in Lymphoma and Lung Cancer by Using Deep Convolutional Neural Networks, Radiology. 294 (2019) 445–452. https://doi.org/10.1148/radiol.2019191114.

[28] S. Röhrich, J. Hofmanninger, F. Prayer, H. Müller, H. Prosch, G. Langs, Prospects and Challenges of Radiomics by Using Nononcologic Routine Chest CT, Radiol. Cardiothorac. Imaging. 2 (2020) e190190. https://doi.org/10.1148/ryct.2020190190.

[29] M. Bogowicz, R.T.H. Leijenaar, S. Tanadini-Lang, O. Riesterer, M. Pruschy, G. Studer, J. Unkelbach, M. Guckenberger, E. Konukoglu, P. Lambin, Post-radiochemotherapy PET radiomics in head and neck cancer – The influence of radiomics implementation on the reproducibility of local control tumor models, Radiother. Oncol. 125 (2017) 385–391. https://doi.org/10.1016/j.radonc.2017.10.023.

[30] J.E. van Timmeren, R.T.H. Leijenaar, W. van Elmpt, B. Reymen, C. Oberije, R. Monshouwer, J. Bussink, C. Brink, O. Hansen, P. Lambin, Survival prediction of non-small cell lung cancer patients using radiomics analyses of cone-beam CT images, Radiother. Oncol. 123 (2017) 363–369. https://doi.org/10.1016/j.radonc.2017.04.016.

[31] S.S.F. Yip, H.J.W.L. Aerts, Applications and limitations of radiomics, Phys. Med. Biol. 61 (2016) R150–R166. https://doi.org/10.1088/0031-9155/61/13/r150.

[32] F. Orlhac, F. Frouin, C. Nioche, N. Ayache, I. Buvat, Validation of a method to compensate multicenter effects affecting CT radiomics, Radiology. 291 (2019) 53–59. https://doi.org/10.1148/radiol.2019182023.

[33] A. Zwanenburg, M. Vallières, M.A. Abdalah, H.J.W.L. Aerts, V. Andrearczyk, A. Apte, S. Ashrafinia, S. Bakas, R.J. Beukinga, R. Boellaard, M. Bogowicz, L. Boldrini, I. Buvat, G.J.R. Cook, C. Davatzikos, A. Depeursinge, M.C. Desseroit, N. Dinapoli, C.V. Dinh, S. Echegaray, The image biomarker standardization initiative: Standardized quantitative radiomics for high-throughput image-based phenotyping, Radiology. 295 (2020) 328–338. https://doi.org/10.1148/radiol.2020191145.

[34] D.F. Polan, S.L. Brady, R.A. Kaufman, Tissue segmentation of computed tomography images using a Random Forest algorithm: a feasibility study, Phys. Med. Biol. 61 (2016) 6553–6569. https://doi.org/10.1088/0031-9155/61/17/6553.

[35] N. Porz, S. Bauer, A. Pica, P. Schucht, J. Beck, R.K. Verma, J. Slotboom, M. Reyes, R. Wiest, Multi-modal glioblastoma segmentation: Man versus machine, PLoS One. 9 (2014) 1–9. https://doi.org/10.1371/journal.pone.0096873.

[36]   N. Porz, S. Habegger, R. Meier, R. Verma, A. Jilch, J. Fichtner, U. Knecht, C. Radina, P. Schucht, J. Beck, A. Raabe, J. Slotboom, M. Reyes, R. Wiest, Fully automated enhanced tumor compartmentalization: Man vs. Machine reloaded, PLoS One. 11 (2016) 1–16. https://doi.org/10.1371/journal.pone.0165302.

[37]   Y. Balagurunathan, Y. Gu, H. Wang, V. Kumar, O. Grove, S. Hawkins, J. Kim, D.B. Goldgof, L.O. Hall, R.A. Gatenby, R.J. Gillies, Reproducibility and Prognosis of Quantitative Features Extracted from CT Images, Transl. Oncol. 7 (2014) 72–87. https://doi.org/https://doi.org/10.1593/tlo.13844.

[38]   W. Grootjans, F. Tixier, C.S. van der Vos, D. Vriens, C.C. Le Rest, J. Bussink, W.J.G. Oyen, L.-F. de Geus-Oei, D. Visvikis, E.P. Visser, The Impact of Optimal Respiratory Gating and Image Noise on Evaluation of Intratumor Heterogeneity on 18F-FDG PET Imaging of Lung Cancer., J. Nucl. Med. 57 (2016) 1692–1698. https://doi.org/10.2967/jnumed.116.173112.

[39]   A. Zwanenburg, S. Leger, L. Agolli, K. Pilz, E.G.C. Troost, C. Richter, S. Löck, Assessing robustness of radiomic features by image perturbation, Sci. Rep. 9 (2019) 614. https://doi.org/10.1038/s41598-018-36938-4.

[40]   A.A. Farag, H.E.A.E. Munim, J.H. Graham, A.A. Farag, A Novel Approach for Lung Nodules Segmentation in Chest CT Using Level Sets, IEEE Trans. Image Process. 22 (2013) 5202–5213. https://doi.org/10.1109/TIP.2013.2282899.

[41]   J. Guiot, A. Vaidyanathan, L. Deprez, F. Zerka, D. Danthine, A.-N. Frix, M. Thys, M. Henket, G. Canivet, S. Mathieu, E. Eftaxia, P. Lambin, N. Tsoutzidis, B. Miraglio, S. Walsh, M. Moutschen, R. Louis, P. Meunier, W. Vos, R.T.H. Leijenaar, P. Lovinfosse, Development and Validation of an Automated Radiomic CT Signature for Detecting COVID-19, Diagnostics . 11 (2021). https://doi.org/10.3390/diagnostics11010041.

[42]   D. Mackin, X. Fave, L. Zhang, D. Fried, J. Yang, B. Taylor, E. Rodriguez-Rivera, C. Dodge, A.K. Jones, L. Court, Measuring Computed Tomography Scanner Variability of Radiomics Features, Invest. Radiol. 50 (2015). https://journals.lww.com/investigativeradiology/Fulltext/2015/11000/Measuring_Computed_Tomography_Scanner_Variability.3.aspx.

[43]   R. Berenguer, M. Del Rosario Pastor-Juan, J. Canales-Vázquez, M. Castro-García, M.V. Villas, F.M. Legorburo, S. Sabater, Radiomics of CT features may be nonreproducible and redundant: Influence of CT acquisition parameters, Radiology. 288 (2018) 407–415. https://doi.org/10.1148/radiol.2018172361.

[44]   B. Zhao, Y. Tan, W.-Y. Tsai, J. Qi, C. Xie, L. Lu, L.H. Schwartz, Reproducibility of radiomics for deciphering tumor phenotype with imaging, Sci. Rep. 6 (2016) 23428. https://doi.org/10.1038/srep23428.

[45]   C. Haarburger, G. Müller-Franzes, L. Weninger, C. Kuhl, D. Truhn, D. Merhof, Radiomics feature reproducibility under inter-rater variability in segmentations of CT images, Sci. Rep. 10 (2020) 1–10. https://doi.org/10.1038/s41598-020-69534-6.

[46]   J.E. van Timmeren, R.T.H. Leijenaar, W. van Elmpt, J. Wang, Z. Zhang, A. Dekker, P. Lambin, Test-Retest Data for Radiomics Feature Stability Analysis: Generalizable or Study-Specific?, Tomogr. (Ann Arbor, Mich.). 2 (2016) 361–365. https://doi.org/10.18383/j.tom.2016.00208.

[47]   J. Peerlings, H.C. Woodruff, J.M. Winfield, A. Ibrahim, B.E. Van Beers, A. Heerschap, A. Jackson, J.E. Wildberger, F.M. Mottaghy, N.M. DeSouza, P. Lambin, Stability of radiomics features in apparent diffusion coefficient maps from a multi-centre test-retest trial, Sci. Rep. 9 (2019) 4800. https://doi.org/10.1038/s41598-019-41344-5.

[48]   R. Da-ano, I. Masson, F. Lucia, M. Doré, P. Robin, J. Alfieri, C. Rousseau, A. Mervoyer, C. Reinhold, J. Castelli, R. De Crevoisier, J.F. Rameé, O. Pradier, U. Schick, D. Visvikis, M. Hatt, Performance comparison of modified ComBat for harmonization of radiomic features for multicenter studies, Sci. Rep. 10 (2020) 1–12. https://doi.org/10.1038/s41598-020-66110-w.

[49]   F. Orlhac, S. Boughdad, C. Philippe, H. Stalla-Bourdillon, C. Nioche, L. Champion, M. Soussan, F. Frouin, V. Frouin, I. Buvat, A postreconstruction harmonization method for multicenter radiomic studies in PET, J. Nucl. Med. 59 (2018) 1321–1328. https://doi.org/10.2967/jnumed.117.199935.

[50]   V. Andrearczyk, A. Depeursinge, H. Müller, Neural network training for cross-protocol radiomic feature standardization in computed tomography, J. Med. Imaging. 6 (2019) 1. https://doi.org/10.1117/1.jmi.6.2.024008.

[51]   A. Crombé, M. Kind, D. Fadli, F. Le Loarer, A. Italiano, X. Buy, O. Saut, Intensity harmonization techniques influence radiomics features and radiomics-based predictions in sarcoma patients, Sci. Rep. 10 (2020) 1–13. https://doi.org/10.1038/s41598-020-72535-0.

[52] V. Sandfort, K. Yan, P.J. Pickhardt, R.M. Summers, Data augmentation using generative adversarial networks (CycleGAN) to improve generalizability in CT segmentation tasks, Sci. Rep. 9 (2019) 1–9. https://doi.org/10.1038/s41598-019-52737-x.

[53] R. Da-Ano, D. Visvikis, M. Hatt, Harmonization strategies for multicenter radiomics investigations, Phys. Med. Biol. 65 (2020) 24TR02. https://doi.org/10.1088/1361-6560/aba798.

[54] S. Ranjbar, J. Ross Mitchell, Chapter 8 - An Introduction to Radiomics: An Evolving Cornerstone of Precision Medicine, in: A. Depeursinge, O. S. Al-Kadi, J.R.B.T.-B.T.A. Mitchell (Eds.), Academic Press, 2017: pp. 223–245. https://doi.org/https://doi.org/10.1016/B978-0-12-812133-7.00008-9.

[55] L. Antonelli, M.R. Guarracino, L. Maddalena, M. Sangiovanni, Integrating imaging and omics data: A review, Biomed. Signal Process. Control. 52 (2019) 264–280. https://doi.org/https://doi.org/10.1016/j.bspc.2019.04.032.

[56] M. Hatt, F. Tixier, L. Pierce, P.E. Kinahan, C.C. Le Rest, D. Visvikis, Characterization of PET/CT images using texture analysis: the past, the present… any future?, Eur. J. Nucl. Med. Mol. Imaging. 44 (2017) 151–165. https://doi.org/10.1007/s00259-016-3427-0.

[57] Y.-H.D. Fang, C.-Y. Lin, M.-J. Shih, H.-M. Wang, T.-Y. Ho, C.-T. Liao, T.-C. Yen, Development and evaluation of an open-source software package "CGITA" for quantifying tumor heterogeneity with molecular images., Biomed Res. Int. 2014 (2014) 248505. https://doi.org/10.1155/2014/248505.

[58] L. Zhang, D V Fried, X.J. Fave, L.A. Hunter, J. Yang, L.E. Court, IBEX: an open infrastructure software platform to facilitate collaborative work in radiomics., Med. Phys. 42 (2015) 1341–1353. https://doi.org/10.1118/1.4908210.

[59] D.A.P. Delzell, S. Magnuson, T. Peter, M. Smith, B.J. Smith, Machine Learning and Feature Selection Methods for Disease Classification With Application to Lung Cancer Screening Image Data, Front. Oncol. 9 (2019) 1393. https://www.frontiersin.org/article/10.3389/fonc.2019.01393.

[60] M. Radovic, M. Ghalwash, N. Filipovic, Z. Obradovic, Minimum redundancy maximum relevance feature selection approach for temporal gene expression data, BMC Bioinformatics. 18 (2017) 9. https://doi.org/10.1186/s12859-016-1423-9.

[61] X. Ying, An Overview of Overfitting and its Solutions, (2019) 22022. https://doi.org/10.1088/1742-6596/1168/2/022022.

[62] H. Song, M. Kim, D. Park, J.-G. Lee, PRESTOPPING: HOW DOES EARLY STOPPING HELP GENERALIZATION AGAINST LABEL NOISE?, n.d.

[63] S.-H. Lee, H. Cho, H.Y. Lee, H. Park, Clinical impact of variability on CT radiomics and suggestions for suitable feature selection: a focus on lung cancer, Cancer Imaging. 19 (2019) 54. https://doi.org/10.1186/s40644-019-0239-z.

[64] Q. Qiu, J. Duan, Z. Duan, X. Meng, C. Ma, J. Zhu, J. Lu, T. Liu, Y. Yin, Reproducibility and non-redundancy of radiomic features extracted from arterial phase CT scans in hepatocellular carcinoma patients: impact of tumor segmentation variability, Quant. Imaging Med. Surgery; Vol 9, No 3 (March 2019) Quant. Imaging Med. Surg. (2019). http://qims.amegroups.com/article/view/24564.

[65] C. Parmar, P. Grossmann, J. Bussink, P. Lambin, H.J.W.L. Aerts, Machine Learning methods for Quantitative Radiomic Biomarkers, Sci. Rep. 5 (2015) 13087. https://doi.org/10.1038/srep13087.

[66] N. Garau, C. Paganelli, P. Summers, W. Choi, S. Alam, W. Lu, C. Fanciullo, M. Bellomi, G. Baroni, C. Rampinelli, External validation of radiomics-based predictive models in low-dose CT screening for early lung cancer diagnosis, Med. Phys. n/a (2020). https://doi.org/10.1002/mp.14308.

[67] T.P.A. Debray, Y. Vergouwe, H. Koffijberg, D. Nieboer, E.W. Steyerberg, K.G.M. Moons, A new framework to enhance the interpretation of external validation studies of clinical prediction models, J. Clin. Epidemiol. 68 (2015) 279–289. https://doi.org/10.1016/j.jclinepi.2014.06.018.

[68] J.E. Park, S.Y. Park, H.J. Kim, H.S. Kim, Reproducibility and Generalizability in Radiomics Modeling: Possible Strategies in Radiologic and Statistical Perspectives., Korean J. Radiol. 20 (2019) 1124–1137. https://doi.org/10.3348/kjr.2018.0070.

[69] N. Altman, M. Krzywinski, The curse(s) of dimensionality, Nat. Methods. 15 (2018) 399–400. https://doi.org/10.1038/s41592-018-0019-x.

[70] E. Capobianco, J. Deng, Radiomics at a Glance: A Few Lessons Learned from Learning Approaches, Cancers . 12 (2020). https://doi.org/10.3390/cancers12092453.

[71] M. Avanzo, L. Wei, J. Stancanello, M. Vallières, A. Rao, O. Morin, S.A. Mattonen, I. El Naqa, Machine and deep learning methods for radiomics, Med. Phys. 47 (2020) e185–e202. https://doi.org/10.1002/mp.13678.

2

[72]   Q.-P. Liu, X. Xu, F.-P. Zhu, Y.-D. Zhang, X.-S. Liu, Prediction of prognostic risk factors in hepatocellular carcinoma with transarterial chemoembolization using multi-modal multi-task deep learning, EClinicalMedicine. 23 (2020). https://doi.org/10.1016/j.eclinm.2020.100379.

[73]   P. Giraud, P. Giraud, A. Gasnier, R. El Ayachy, S. Kreps, J.-P. Foy, C. Durdux, F. Huguet, A. Burgun, J.-E. Bibault, Radiomics and Machine Learning for Radiotherapy in Head and Neck Cancers , Front. Oncol. 9 (2019) 174. https://www.frontiersin.org/article/10.3389/fonc.2019.00174.

[74]   G. Choy, O. Khalilzadeh, M. Michalski, S. Do, A.E. Samir, O.S. Pianykh, J.R. Geis, P. V Pandharipande, J.A. Brink, K.J. Dreyer, Current Applications and Future Impact of Machine Learning in Radiology, Radiology. 288 (2018) 318–328. https://doi.org/10.1148/radiol.2018171820.

[75]   https://github.com/, (n.d.).

[76]   https://about.gitlab.com/, (n.d.).

[77]   https://www.openml.org/, (n.d.).

[78]   E.R. DeLong, D.M. DeLong, D.L. Clarke-Pearson, Comparing the Areas under Two or More Correlated Receiver Operating Characteristic Curves: A Nonparametric Approach, Biometrics. 44 (1988) 837–845. https://doi.org/10.2307/2531595.

[79]   E.W. Steyerberg, A.J. Vickers, N.R. Cook, T. Gerds, M. Gonen, N. Obuchowski, M.J. Pencina, M.W. Kattan, Assessing the performance of prediction models: a framework for traditional and novel measures., Epidemiology. 21 (2010) 128–138. https://doi.org/10.1097/EDE.0b013e3181c30fb2.

[80]   G.S. Collins, J.B. Reitsma, D.G. Altman, K.G.M. Moons, Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): the TRIPOD Statement, BMC Med. 13 (2015) 1. https://doi.org/10.1186/s12916-014-0241-z.

[81]   S. Lemeshow, D.W. Hosmer Jr., A review of goodness of fit statistics for use in the development of logistic regression models1, Am. J. Epidemiol. 115 (1982) 92–106. https://doi.org/10.1093/oxfordjournals.aje.a113284.

[82]   J.T. Leek, R.D. Peng, Statistics: P values are just the tip of the iceberg., Nature. 520 (2015) 612. https://doi.org/10.1038/520612a.

[83]   C. Drummond, Replicability is not reproducibility: nor is it good science, Proc. Eval. Methods Mach. Learn. Work. 26th ICML, Montr. Canada,2009. (2009).

[84]   R.D. Peng, Reproducible Research in Computational Science, Science (80-.). 334 (2011) 1226 LP – 1227. https://doi.org/10.1126/science.1213847.

[85]   R.D. Peng, F. Dominici, S.L. Zeger, Reproducible Epidemiologic Research, Am. J. Epidemiol. 163 (2006) 783–789. https://doi.org/10.1093/aje/kwj093.

[86]   S. Fiset, M.L. Welch, J. Weiss, M. Pintilie, J.L. Conway, M. Milosevic, A. Fyles, A. Traverso, D. Jaffray, U. Metser, J. Xie, K. Han, Repeatability and reproducibility of MRI-based radiomic features in cervical cancer, Radiother. Oncol. 135 (2019) 107–114. https://doi.org/10.1016/j.radonc.2019.03.001.

[87]   P. Kalendralis, Z. Shi, A. Traverso, A. Choudhury, M. Sloep, I. Zhovannik, M.P.A. Starmans, D. Grittner, P. Feltens, R. Monshouwer, S. Klein, R. Fijten, H. Aerts, A. Dekker, J. van Soest, L. Wee, FAIR-compliant clinical, radiomics and DICOM metadata of RIDER, interobserver, Lung1 and head-Neck1 TCIA collections, Med. Phys. 47 (2020) 5931–5940. https://doi.org/https://doi.org/10.1002/mp.14322.

[88]   F. Valdora, N. Houssami, F. Rossi, M. Calabrese, A.S. Tagliafico, Rapid review: radiomics and breast cancer, Breast Cancer Res. Treat. 169 (2018) 217–229. https://doi.org/10.1007/s10549-018-4675-4.

[89]   R.W.Y. Granzier, T.J.A. van Nijnatten, H.C. Woodruff, M.L. Smidt, M.B.I. Lobbes, Exploring breast cancer response prediction to neoadjuvant systemic therapy using MRI-based radiomics: A systematic review, Eur. J. Radiol. 121 (2019). https://doi.org/10.1016/j.ejrad.2019.108736.

[90]   T. Wakabayashi, F. Ouhmich, C. Gonzalez-Cabrera, E. Felli, A. Saviano, V. Agnus, P. Savadjiev, T.F. Baumert, P. Pessaux, J. Marescaux, B. Gallix, Radiomics in hepatocellular carcinoma: a quantitative review, Hepatol. Int. 13 (2019) 546–559. https://doi.org/10.1007/s12072-019-09973-0.

[91]   J.E. Park, H.S. Kim, D. Kim, S.Y. Park, J.Y. Kim, S.J. Cho, J.H. Kim, A systematic review reporting quality of radiomics research in neuro-oncology: Toward clinical utility and quality improvement using high-dimensional imaging features, BMC Cancer. 20 (2020) 1–11. https://doi.org/10.1186/s12885-019-6504-5.

[92]   A. Stanzione, M. Gambardella, R. Cuocolo, A. Ponsiglione, V. Romeo, M. Imbriaco, Prostate MRI radiomics: A systematic review and radiomic quality score assessment, Eur. J. Radiol. 129 (2020) 109095. https://doi.org/10.1016/j.ejrad.2020.109095.

[93] M.R. Chetan, F. V. Gleeson, Radiomics in predicting treatment response in non-small-cell lung cancer: current status, challenges and future perspectives, Eur. Radiol. (2020). https://doi.org/10.1007/s00330-020-07141-9.

[94] M. Abdurixiti, M. Nijiati, R. Shen, Q. Ya, N. Abuduxiku, M. Nijiati, Current progress and quality of radiomic studies for predicting EGFR mutation in patients with non-small cell lung cancer using PET/CT images: a systematic review., Br. J. Radiol. (2021) 20201272. https://doi.org/10.1259/bjr.20201272.

[95] S. Ursprung, L. Beer, A. Bruining, R. Woitek, G.D. Stewart, F.A. Gallagher, E. Sala, Radiomics of computed tomography and magnetic resonance imaging in renal cell carcinoma—a systematic review and meta-analysis, Eur. Radiol. 30 (2020) 3558–3566. https://doi.org/10.1007/s00330-020-06666-3.

[96] G.L. Banna, T. Olivier, F. Rundo, U. Malapelle, F. Fraggetta, M. Libra, A. Addeo, The Promise of Digital Biopsy for the Prediction of Tumor Molecular Features and Clinical Outcomes Associated With Immunotherapy, Front. Med. 6 (2019) 172. https://doi.org/10.3389/fmed.2019.00172.

[97] P. Martin-Gonzalez, M. Crispin-Ortuzar, L. Rundo, M. Delgado-Ortet, M. Reinius, L. Beer, R. Woitek, S. Ursprung, H. Addley, J.D. Brenton, F. Markowetz, E. Sala, Integrative radiogenomics for virtual biopsy and treatment monitoring in ovarian cancer, Insights Imaging. 11 (2020). https://doi.org/10.1186/s13244-020-00895-2.

[98] B. Shofty, M. Artzi, S. Shtrozberg, C. Fanizzi, F. DiMeco, O. Haim, S. Peleg Hason, Z. Ram, D. Ben Bashat, R. Grossman, Virtual biopsy using MRI radiomics for prediction of BRAF status in melanoma brain metastasis, Sci. Rep. 10 (2020) 1–7. https://doi.org/10.1038/s41598-020-63821-y.

[99] E.J. Limkin, R. Sun, Radiomics to predict response to immunotherapy: an imminent reality?, Future Oncol. 16 (2020) 1673–1676. https://doi.org/10.2217/fon-2020-0015.

[100] L. Tselikas, R. Sun, S. Ammari, L. Dercle, S. Yevich, A. Hollebecque, M. Ngo-Camus, C. Nicotra, E. Deutsch, F. Deschamps, T. de Baere, Role of image-guided biopsy and radiomics in the age of precision medicine, Chinese Clin. Oncol. 8 (2019) 6–13. https://doi.org/10.21037/cco.2019.12.02.

[101] M. Ismail, R. Correa, K. Bera, R. Verma, A.S. Bamashmos, N. Beig, J. Antunes, P. Prasanna, V. Statsevych, M. Ahluwalia, P. Tiwari, Spatial-And-Context aware (SpACe) "virtual biopsy" radiogenomic maps to target tumor mutational status on structural MRI, (2020) 1–10. http://arxiv.org/abs/2006.09878.

[102] S. Ranjbar, S.N. Velgos, A.C. Dueck, Y.E. Geda, J.R. Mitchell, Brain MR radiomics to differentiate cognitive disorders, J. Neuropsychiatry Clin. Neurosci. 31 (2019) 210–219. https://doi.org/10.1176/appi.neuropsych.17120366.

[103] H. Zhou, J. Jiang, J. Lu, M. Wang, H. Zhang, C. Zuo, Dual-model radiomic biomarkers predict development of mild cognitive impairment progression to Alzheimer's disease, Front. Neurosci. 13 (2019) 1–13. https://doi.org/10.3389/fnins.2018.01045.

[104] K. Zhao, Y. Ding, Y. Han, Y. Fan, A.F. Alexander-Bloch, T. Han, D. Jin, B. Liu, J. Lu, C. Song, P. Wang, D. Wang, Q. Wang, K. Xu, H. Yang, H. Yao, Y. Zheng, C. Yu, B. Zhou, X. Zhang, Y. Zhou, T. Jiang, X. Zhang, Y. Liu, Independent and reproducible hippocampal radiomic biomarkers for multisite Alzheimer's disease: diagnosis, longitudinal progress and biological basis, Sci. Bull. 65 (2020) 1103–1113. https://doi.org/10.1016/j.scib.2020.04.003.

[105] A. Chaddad, C. Desrosiers, L. Hassan, C. Tanougast, Hippocampus and amygdala radiomic biomarkers for the study of autism spectrum disorder, BMC Neurosci. 18 (2017) 1–12. https://doi.org/10.1186/s12868-017-0373-0.

[106] Q. Feng, Q. Song, M. Wang, P.P. Pang, Z. Liao, H. Jiang, D. Shen, Z. Ding, Hippocampus Radiomic Biomarkers for the Diagnosis of Amnestic Mild Cognitive Impairment: A Machine Learning Method, Front. Aging Neurosci. 11 (2019) 1–10. https://doi.org/10.3389/fnagi.2019.00323.

[107] Y. Li, J. Jiang, J. Lu, J. Jiang, H. Zhang, C. Zuo, Radiomics: a novel feature extraction method for brain neuron degeneration disease using 18F-FDG PET imaging and its implementation for Alzheimer's disease and mild cognitive impairment, Ther. Adv. Neurol. Disord. 12 (2019) 1–21. https://doi.org/10.1177/1756286419838682.

[108] L. He, Z. Liu, C. Liu, Z. Gao, Q. Ren, L. Lei, J. Ren, Radiomics Based on Lumbar Spine Magnetic Resonance Imaging to Detect Osteoporosis, Acad. Radiol. (2020) 1–7. https://doi.org/10.1016/j.acra.2020.03.046.

[109] S. Rastegar, M. Vaziri, Y. Qasempour, M.R. Akhash, N. Abdalvand, I. Shiri, H. Abdollahi, H. Zaidi, Radiomics for classification of bone mineral loss: A machine learning study, Diagn. Interv. Imaging. 101 (2020) 599–610. https://doi.org/10.1016/j.diii.2020.01.008.

2

[110] J. Bianchi, J.R. Gonçalves, A.C. De Oliveira Ruellas, J.B. Vimort, M. Yatabe, B. Paniagua, P. Hernandez, E. Benavides, F.N. Soki, L.H.S. Cevidanes, Software comparison to analyze bone radiomics from high resolution CBCT scans of mandibular condyles, Dentomaxillofacial Radiol. 48 (2019) 1–10. https://doi.org/10.1259/dmfr.20190049.

[111] H. Wang, R. Wei, G. Rao, J. Zhu, B. Song, Characteristic CT findings distinguishing 2019 novel coronavirus disease (COVID-19) from influenza pneumonia, Eur. Radiol. 30 (2020) 4910–4917. https://doi.org/10.1007/s00330-020-06880-z.

[112] M.-P. Revel, A.P. Parkar, H. Prosch, M. Silva, N. Sverzellati, F. Gleeson, A. Brady, COVID-19 patients and the radiology department – advice from the European Society of Radiology (ESR) and the European Society of Thoracic Imaging (ESTI), Eur. Radiol. (2020) 1–7. https://doi.org/10.1007/s00330-020-06865-y.

[113] J. Guiot, A. Vaidyanathan, L. Deprez, F. Zerka, D. Danthine, A.-N. Frix, M. Thys, M. Henket, G. Canivet, S. Mathieu, E. Eftaxia, P. Lambin, N. Tsoutzidis, B. Miraglio, S. Walsh, M. Moutschen, R. Louis, P. Meunier, W. Vos, R. Leijenaar, P. Lovinfosse, Development and validation of an automated radiomic CT signature for detecting COVID-19, MedRxiv. (2020) 2020.04.28.20082966. https://doi.org/10.1101/2020.04.28.20082966.

[114] L. Li, L. Qin, Z. Xu, Y. Yin, X. Wang, B. Kong, J. Bai, Y. Lu, Z. Fang, Q. Song, K. Cao, D. Liu, G. Wang, Q. Xu, X. Fang, S. Zhang, J. Xia, J. Xia, Using Artificial Intelligence to Detect COVID-19 and Community-acquired Pneumonia Based on Pulmonary CT: Evaluation of the Diagnostic Accuracy, Radiology. 296 (2020) E65–E71. https://doi.org/10.1148/radiol.2020200905.

[115] Q. Zeng, K.I. Zheng, J. Chen, Z. Jiang, T. Tian, X. Wang, H. Ma, K. Pan, Y. Yang, Y. Chen, M. Zheng, Radiomics-based model for accurately distinguishing between severe acute respiratory syndrome associated coronavirus 2 (SARS-CoV-2) and influenza A infected pneumonia, MedComm. (2020) 1–9. https://doi.org/10.1002/mco2.14.

[116] Y. Chen, Y. Wang, Y. Zhang, N. Zhang, S. Zhao, H. Zeng, W. Deng, Z. Huang, S. Liu, B. Song, A quantitative and radiomics approach to monitoring ards in COVID-19 patients based on chest CT: A retrospective cohort study, Int. J. Med. Sci. 17 (2020) 1773–1782. https://doi.org/10.7150/ijms.48432.

[117] H.X. Bai, R. Wang, Z. Xiong, B. Hsieh, K. Chang, K. Halsey, T.M.L. Tran, J.W. Choi, D.-C. Wang, L.-B. Shi, J. Mei, X.-L. Jiang, I. Pan, Q.-H. Zeng, P.-F. Hu, Y.-H. Li, F.-X. Fu, R.Y. Huang, R. Sebro, Q.-Z. Yu, M.K. Atalay, W.-H. Liao, Artificial Intelligence Augmentation of Radiologist Performance in Distinguishing COVID-19 from Pneumonia of Other Origin at Chest CT, Radiology. 296 (2020) E156–E165. https://doi.org/10.1148/radiol.2020201491.

[118] H. Yue, Q. Yu, C. Liu, Y. Huang, Z. Jiang, C. Shao, H. Zhang, B. Ma, Y. Wang, G. Xie, H. Zhang, X. Li, N. Kang, X. Meng, S. Huang, D. Xu, J. Lei, H. Huang, J. Yang, J. Ji, H. Pan, S. Zou, S. Ju, X. Qi, Machine learning-based CT radiomics method for predicting hospital stay in patients with pneumonia associated with SARS-CoV-2 infection: a multicenter study, Ann. Transl. Med. 8 (2020) 859. https://doi.org/10.21037/atm-20-3026.

[119] F. Shi, J. Wang, J. Shi, Z. Wu, Q. Wang, Z. Tang, K. He, Y. Shi, D. Shen, Review of Artificial Intelligence Techniques in Imaging Data Acquisition, Segmentation and Diagnosis for COVID-19, IEEE Rev. Biomed. Eng. (2020) 1. https://doi.org/10.1109/RBME.2020.2987975.

[120] S. Kundu, H. Elhalawani, J.W. Gichoya, C.E. Kahn, How Might AI and Chest Imaging Help Unravel COVID-19's Mysteries?, Radiol. Artif. Intell. 2 (2020) e200053. https://doi.org/10.1148/ryai.2020200053.

[121] X. Fave, L. Zhang, J. Yang, D. Mackin, P. Balter, D. Gomez, D. Followill, A.K. Jones, F. Stingo, Z. Liao, R. Mohan, L. Court, Delta-radiomics features for the prediction of patient outcomes in non-small cell lung cancer., Sci. Rep. 7 (2017) 588. https://doi.org/10.1038/s41598-017-00665-z.

[122] Y. Ma, W. Ma, X. Xu, F. Cao, How Does the Delta-Radiomics Better Differentiate Pre-Invasive GGNs From Invasive GGNs?, Front. Oncol. 10 (2020) 1–7. https://doi.org/10.3389/fonc.2020.01017.

[123] H. Chen, L. Shi, K.N.B. Nguyen, A.M. Monjazeb, K.E. Matsukuma, T.W. Loehfelm, H. Huang, J. Qiu, Y. Rong, MRI Radiomics for Prediction of Tumor Response and Downstaging in Rectal Cancer Patients after Preoperative Chemoradiation, Adv. Radiat. Oncol. (2020) 1–10. https://doi.org/10.1016/j.adro.2020.04.016.

[124] H. Nasief, C. Zheng, D. Schott, W. Hall, S. Tsai, B. Erickson, X. Allen Li, A machine learning based delta-radiomics process for early prediction of treatment response of pancreatic cancer, Npj Precis. Oncol. 3 (2019) 1–10. https://doi.org/10.1038/s41698-019-0096-z.

[125] P. Lin, P.F. Yang, S. Chen, Y.Y. Shao, L. Xu, Y. Wu, W. Teng, X.Z. Zhou, B.H. Li, C. Luo, L.M. Xu, M. Huang, T.Y. Niu, Z.M. Ye, A Delta-radiomics model for preoperative evaluation of Neoadjuvant chemotherapy response in high-grade osteosarcoma, Cancer Imaging. 20 (2020) 1–12. https://doi.org/10.1186/s40644-019-0283-8.

[126] Y. Liu, M. Wu, Y. Zhang, Y. Luo, S. He, Y. Wang, F. Chen, Y. Liu, Q. Yang, Y. Li, H. Wei, H. Zhang, C. Jin, N. Lu, W. Li, S. Wang, Y. Guo, Z. Ye, Imaging Biomarkers to Predict and Evaluate the Effectiveness of Immunotherapy in Advanced Non-Small-Cell Lung Cancer, Front. Oncol. 11 (2021) 773. https://doi.org/10.3389/fonc.2021.657615.

[127] R.R. Colen, C. Rolfo, M. Ak, M. Ayoub, S. Ahmed, N. Elshafeey, P. Mamindla, P.O. Zinn, C. Ng, R. Vikram, S. Bakas, C.B. Peterson, J. Rodon Ahnert, V. Subbiah, D.D. Karp, B. Stephen, J. Hajjar, A. Naing, Radiomics analysis for predicting pembrolizumab response in patients with advanced rare cancers, J. Immunother. Cancer. 9 (2021) e001752. https://doi.org/10.1136/jitc-2020-001752.

[128] K. Fatima, A. Dasgupta, D. DiCenzo, C. Kolios, K. Quiaoit, M. Saifuddin, M. Sandhu, D. Bhardwaj, I. Karam, I. Poon, Z. Husain, L. Sannachi, G.J. Czarnota, Ultrasound delta-radiomics during radiotherapy to predict recurrence in patients with head and neck squamous cell carcinoma, Clin. Transl. Radiat. Oncol. 28 (2021) 62–70. https://doi.org/10.1016/j.ctro.2021.03.002.

[129] J.O. Deasy, S.M. Bentzen, A. Jackson, R.K. Ten Haken, E.D. Yorke, L.S. Constine, A. Sharma, L.B. Marks, Improving normal tissue complication probability models: the need to adopt a "data-pooling" culture., Int. J. Radiat. Oncol. Biol. Phys. 76 (2010) S151-4. https://doi.org/10.1016/j.ijrobp.2009.06.094.

[130] T. Skripcak, C. Belka, W. Bosch, C. Brink, T. Brunner, V. Budach, D. Büttner, J. Debus, A. Dekker, C. Grau, S. Gulliford, C. Hurkmans, U. Just, M. Krause, P. Lambin, J.A. Langendijk, R. Lewensohn, A. Lühr, P. Maingon, M. Masucci, M. Niyazi, P. Poortmans, M. Simon, H. Schmidberger, E. Spezi, M. Stuschke, V. Valentini, M. Verheij, G. Whitfield, B. Zackrisson, D. Zips, M. Baumann, Creating a data exchange strategy for radiotherapy research: towards federated databases and anonymised public datasets., Radiother. Oncol. J. Eur. Soc. Ther. Radiol. Oncol. 113 (2014) 303–309. https://doi.org/10.1016/j.radonc.2014.10.001.

[131] R.L. Schilsky, D.L. Michels, A.H. Kearbey, P.P. Yu, C.A. Hudis, Building a rapid learning health care system for oncology: the regulatory framework of CancerLinQ., J. Clin. Oncol. Off. J. Am. Soc. Clin. Oncol. 32 (2014) 2373–2379. https://doi.org/10.1200/JCO.2014.56.2124.

[132] http://www.eurocat.info/, (n.d.).

[133] T.M. Deist, A. Jochems, J. van Soest, G. Nalbantov, C. Oberije, S. Walsh, M. Eble, P. Bulens, P. Coucke, W. Dries, A. Dekker, P. Lambin, Infrastructure and distributed learning methodology for privacy-preserving multi-centric rapid learning health care: euroCAT, Clin. Transl. Radiat. Oncol. 4 (2017) 24–31. https://doi.org/https://doi.org/10.1016/j.ctro.2016.12.004.

[134] P. Lambin, E. Roelofs, B. Reymen, E.R. Velazquez, J. Buijsen, C.M.L. Zegers, S. Carvalho, R.T.H. Leijenaar, G. Nalbantov, C. Oberije, M. Scott Marshall, F. Hoebers, E.G.C. Troost, R.G.P.M. van Stiphout, W. van Elmpt, T. van der Weijden, L. Boersma, V. Valentini, A. Dekker, 'Rapid Learning health care in oncology' – An approach towards decision support systems enabling customised radiotherapy', Radiother. Oncol. 109 (2013) 159–164. https://doi.org/https://doi.org/10.1016/j.radonc.2013.07.007.

[135] The Cancer Imaging Archive, TCIA Collections. cancerimagingarchive.net, (2017). https://www.cancerimagingarchive.net/.

[136] National Cancer Institute, Quantitative Imaging Network (QIN), (2017). https://imaging.cancer.gov/programs_ resources/specialized_initiatives/qin.htm.

[137] Radiological Society of North America, Quantitative Imaging Biomarkers Alliance® (QIBA®), (2017). https://www.rsna.org/qiba/.

[138] www.medomics.ai, (n.d.).

[139] QuiC ConCePT, (2017). http://www.quic-concept.eu/.

[140] F. Zerka, V. Urovi, A. Vaidyanathan, S. Barakat, R.T.H. Leijenaar, S. Walsh, H. Gabrani-Juma, B. Miraglio, H.C. Woodruff, M. Dumontier, P. Lambin, Blockchain for Privacy Preserving and Trustworthy Distributed Machine Learning in Multicentric Medical Imaging (C-DistriM), IEEE Access. 8 (2020) 183939–183951. https://doi.org/10.1109/ACCESS.2020.3029445.

[141] A. Jochems, T.M. Deist, I. El Naqa, M. Kessler, C. Mayo, J. Reeves, S. Jolly, M. Matuszak, R. Ten Haken, J. van Soest, C. Oberije, C. Faivre-Finn, G. Price, D. de Ruysscher, P. Lambin, A. Dekker, Developing and Validating a Survival Prediction Model for NSCLC Patients Through Distributed Learning Across 3 Countries, Int. J. Radiat. Oncol. Biol. Phys. 99 (2017) 344–352. https://doi.org/10.1016/j.ijrobp.2017.04.021.

2

[142] Z. Shi, I. Zhovannik, A. Traverso, F.J.W.M. Dankers, T.M. Deist, P. Kalendralis, R. Monshouwer, J. Bussink, R. Fijten, H.J.W.L. Aerts, A. Dekker, L. Wee, Distributed radiomics as a signature validation study using the Personal Health Train infrastructure, Sci. Data. 6 (2019) 218. https://doi.org/10.1038/s41597-019-0241-0.

[143] M. Bogowicz, A. Jochems, T.M. Deist, S. Tanadini-Lang, S.H. Huang, B. Chan, J.N. Waldron, S. Bratman, B. O'Sullivan, O. Riesterer, G. Studer, J. Unkelbach, S. Barakat, R.H. Brakenhoff, I. Nauta, S.E. Gazzani, G. Calareso, K. Scheckenbach, F. Hoebers, F.W.R. Wesseling, S. Keek, S. Sanduleanu, R.T.H. Leijenaar, M.R. Vergeer, C.R. Leemans, C.H.J. Terhaard, M.W.M. van den Brekel, O. Hamming-Vrieze, M.A. van der Heijden, H.M. Elhalawani, C.D. Fuller, M. Guckenberger, P. Lambin, Privacy-preserving distributed learning of radiomics to predict overall survival and HPV status in head and neck cancer, Sci. Rep. 10 (2020) 1–10. https://doi.org/10.1038/s41598-020-61297-4.

[144] Z. Shi, A. Traverso, J. van Soest, A. Dekker, L. Wee, Technical Note: Ontology-guided radiomics analysis workflow (O-RAW), Med. Phys. 46 (2019) 5677–5684. https://doi.org/https://doi.org/10.1002/mp.13844.

[145] https://bioportal.bioontology.org/ontologies/RO, (n.d.).

[146] E. Pfaehler, A. Zwanenburg, J.R. de Jong, R. Boellaard, RACAT: An open source and easy to use radiomics calculator tool, PLoS One. 14 (2019) 1–26. https://doi.org/10.1371/journal.pone.0212223.

[147] C. Nioche, F. Orlhac, S. Boughdad, S. Reuzé, J. Goya-Outi, C. Robert, C. Pellot-Barakat, M. Soussan, F. Frouin, I. Buvat, LIFEx: A Freeware for Radiomic Feature Calculation in Multimodality Imaging to Accelerate Advances in the Characterization of Tumor Heterogeneity, Cancer Res. 78 (2018) 4786 LP – 4789. https://doi.org/10.1158/0008-5472.CAN-18-0125.

[148] A. Fedorov, R. Beichel, J. Kalpathy-Cramer, J. Finet, J.-C. Fillion-Robin, S. Pujol, C. Bauer, D. Jennings, F. Fennessy, M. Sonka, J. Buatti, S. Aylward, J. V Miller, S. Pieper, R. Kikinis, 3D Slicer as an image computing platform for the Quantitative Imaging Network, Magn. Reson. Imaging. 30 (2012) 1323–1341. https://doi.org/10.1016/j.mri.2012.05.001.

[149] R. Casale, E. Lavrova, S. Sanduleanu, H.C. Woodruff, P. Lambin, Development and external validation of a non-invasive molecular status predictor of chromosome 1p/19q co-deletion based on MRI radiomics analysis of Low Grade Glioma patients, Eur. J. Radiol. 139 (2021) 109678. https://doi.org/10.1016/j.ejrad.2021.109678.

[150] W. Mu, E. Katsoulakis, C.J. Whelan, K.L. Gage, M.B. Schabath, R.J. Gillies, Radiomics predicts risk of cachexia in advanced NSCLC patients treated with immune checkpoint inhibitors, Br. J. Cancer. (2021). https://doi.org/10.1038/s41416-021-01375-0.

[151] V. Chiappa, M. Interlenghi, C. Salvatore, F. Bertolina, G. Bogani, A. Ditto, F. Martinelli, I. Castiglioni, F. Raspagliesi, Using rADioMIcs and machine learning with ultrasonography for the differential diagnosis of myometRiAL tumors (the ADMIRAL pilot study). Radiomics and differential diagnosis of myometrial tumors, Gynecol. Oncol. (2021). https://doi.org/10.1016/j.ygyno.2021.04.004.

[152] F. Zhang, L.-Z. Zhong, X. Zhao, D. Dong, J.-J. Yao, S.-Y. Wang, Y. Liu, D. Zhu, Y. Wang, G.-J. Wang, Y.-M. Wang, D. Li, J. Wei, J. Tian, H. Shan, A deep-learning-based prognostic nomogram integrating microscopic digital pathology and macroscopic magnetic resonance images in nasopharyngeal carcinoma: a multi-cohort study., Ther. Adv. Med. Oncol. 12 (2020) 1758835920971416. https://doi.org/10.1177/1758835920971416.

[153] C. Alvarez-Jimenez, A.A. Sandino, P. Prasanna, A. Gupta, S.E. Viswanath, E. Romero, Identifying Cross-Scale Associations between Radiomic and Pathomic Signatures of Non-Small Cell Lung Cancer Subtypes: Preliminary Results., Cancers (Basel). 12 (2020). https://doi.org/10.3390/cancers12123663.

[154] C. Geady, H. Keller, I. Siddiqui, J. Bilkey, N.C. Dhani, D.A. Jaffray, Bridging the gap between micro- and macro-scales in medical imaging with textural analysis - A biological basis for CT radiomics classifiers?, Phys. Medica PM an Int. J. Devoted to Appl. Phys. to Med. Biol. Off. J. Ital. Assoc. Biomed. Phys. 72 (2020) 142–151. https://doi.org/10.1016/j.ejmp.2020.03.018.

[155] FDA, Proposed Regulatory Framework for Modifications to Artificial Intelligence / Machine Learning (AI / ML)-Based Software as a Medical Device (SaMD) - Discussion Paper and Request for Feedback, U.S Food Drug Adm. (2019) 1–20.

Distributed Learning

3

Privacy-preserving distributed machine learning from federated databases in healthcare

Fadila Zerka | Samir Barakat | Sean Walsh | Marta Bogowicz |
Ralph T.H. Leijenaar | Arthur Jochems | Benjamin Miraglio |
David Towned | Philippe Lambin

Abstract

Big Data for healthcare is one of the potential solutions to deal with the numerous challenges of healthcare such as rising cost, aging population, precision medicine, universal health coverage and the increase of non-communicable diseases. However, data centralization for big data raises privacy and regulatory concerns.

Covered topics include (1) an introduction to privacy of patient data and distributed learning as a potential solution to preserving this data, a description of the legal context for patient data research, and a definition of machine/deep learning concepts; (2) a presentation of the adopted review protocol; (3) a presentation of the search results; (4) a discussion of the findings, limitations of the review, and future perspectives.

Distributed learning from federated databases, makes data centralization unnecessary. Distributed algorithms iteratively analyze separate databases: essentially sharing research questions and answers between databases instead of the data itself. In other words, one can learn from separate and isolated datasets without patient data ever leaving the individual clinical institutes.

Distributed learning promises great potential to facilitate Big Data for medical application in particular for international consortiums. Our purpose is to review the major implementations of distributed learning in healthcare.

Abstract

Introduction

Law and ethics seek to produce a governance framework for the processing of patient data that produces a solution to the issues that arise between the competing desires of individuals in society for on the one hand privacy and on the other advances in health care. Traditional safeguards to achieve this governance have come from, for example, the anonymization of data, or informed consent. These are not adequate safeguards for the new big data and artificial intelligence methodologies in research: it is increasingly difficult to create anonymous data (rather than pseudonymized/ coded data) or to maintain it against reidentification (through linking of datasets causing accidental or deliberate reidentification). The technology of big data and artificial intelligence (AI), however, itself increasingly offers safeguards to solve the governance problem. In this paper we explore how privacy-preserving distributed machine learning from federated databases might assist governance in healthcare. The paper first outlines the basic parameters of the law and ethics issues, and then discusses machine learning and deep learning. Thereafter, the results of the review are presented and then discussed. The methodology for this paper is that distributed machine learning is an evolving field in computing, with 665 articles published between 2001 and 2018, and the study is based on a literature search and focus on the medical applications of distributed machine learning and provide an up-to-date summary of the field.

The legal context for patient data research

The challenges in law and ethics in relation to Big Data and Artificial Intelligence are well documented and discussed [1-16]. The issue is one of balance: privacy of health data and access to data for research. This issue is likely to become more pronounced with the foreseeable developments in healthcare, notably in relation to rising cost, aging population, precision medicine, universal health coverage and the increase of non-communicable diseases. However, recent developments in laws, for example, in the European Union's General Data Protection Regulation, seem to maintain the traditional approach that seems to favor individualism above solidarity. Individualism is strengthened in the new legislation. There is a narrowing of the definition of informed consent in Article 4.11 of the GDPR, with the unclear inclusion of the necessity for broad consent in scientific research included in Recital 33.

In relation to the continuing ambiguity of the unclear legal landscape for research using and reusing large data sets and linking between data sets, the GDPR is not clear in the area of reidentification of individuals. For the GDPR, part of the problem is clear - when data have the potential when added to other data to identify an individual, then those data are personal data and subject to the Regulation. The question is, is this absolute (any possibility, regardless of remoteness), or is there a reasonableness test? Recital 26 includes such a reasonable test: "To ascertain whether means are reasonably likely to

be used to identify the natural person, account should be taken of all objective factors, such as the costs of and the amount of time required for identification, taking into consideration the available technology at the time of the processing and technological developments."

From this overview of legal difficulties, it is clear that there are obstacles to processing data in Big Data, Machine Learning and Artificial Intelligence methodologies and environments. It must be stressed that the object is not to circumvent the rights of patients, or to suggest that privacy should be ignored. The difficulty is that where the law is unclear, there is a tendency towards restrictive readings of the law to avoid liability, and in the case of the methodologies and applications of data science discussed here, the effect of unclear law and restrictive interpretations of the law will be to block potentially important medical and scientific developments and research. Each of the uncertainties will require regulators to take a position on the best interpretation of the meaning of the law according to the available safeguards. The question for the data science community is, how far can that community itself address concerns about privacy, about reidentification, and about safeguarding autonomy of individuals and their legitimate expectations to dignity in their treatment through the proper treatment of their personal data? How far distributed learning might contribute a suitable safeguard is the question addressed in the remainder of this paper.

Machine learning
Machine learning comes from the possibility to apply algorithms on raw data to acquire knowledge [1]. These algorithms are implemented to support decision-making in different domains including healthcare, manufacturing, education, financial modeling, and marketing [2,3]. In medical disciplines, machine learning has contributed to improving the efficiency of clinical trials and decision-making processes. Some examples of machine learning applications in medicine are the localization of thoracic diseases [4], early diagnosis of Alzheimer [5], personalized treatment [6], outcome prediction [7,8], and automated radiology reports [9].

There are three main categories of machine learning algorithms. Firstly, in supervised learning, the algorithm generates a function for mapping input variables to output variables. In unsupervised learning, the applied algorithms do not have any outcome variable to estimate and the algorithms generate a function mapping for the structure of the data. The third type is referred to as reinforcement learning, whereby in the absence of a training dataset the algorithm trains itself by learning from experiences to make increasingly improved decisions. A reinforcement agent decides what action to perform to accomplish a given task [10,11]. Table 3.1 provides a brief description of selected popular machine learning algorithms across the three categories.

Deep learning

Deep learning is a subset of machine learning which, in turn, is a subset of artificial intelligence [12], as represented in Figure 3.1. The learning process of a deep neural network architecture cascades through multiple nodes in multiple layers, where nodes and layers use the output of the previous nodes and layers as input [13]. The output of a node is calculated by applying an activation function to the weighted average of this node's input. As described by Andrew NG [14], "The analogy to deep learning is that the rocket engine is the deep learning models and the fuel is the huge amounts of data that we can feed into these algorithms." Meaning that the more data are fed into the model the better the performance. Yet, this continuous improvement of the performance in concordance with the amount of the data is not correct for traditional machine learning algorithms reaching a steady performance level that does not improve with the increase of the amount of the training data [15].
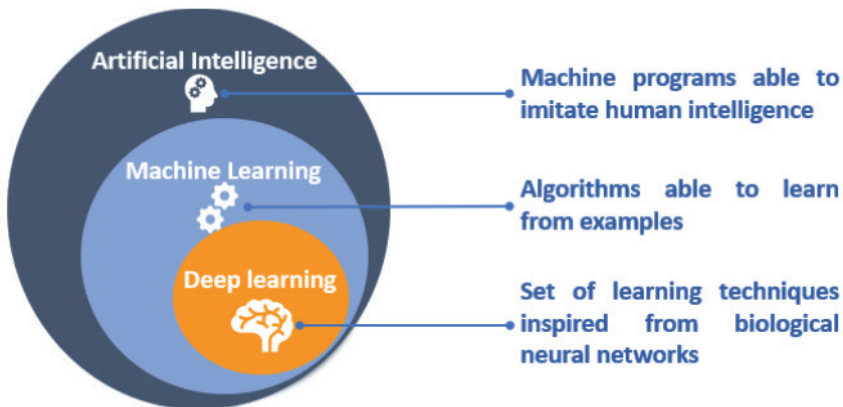

Figure 3.1: Relationship between artificial intelligence, machine learning, and deep learning.

Methods and material selection

A PubMed search was performed to collect relevant studies concerning the utilization of distributed machine learning in medicine. We used the search strings: "distributed learning", "Distributed machine learning" and "privacy preserving data mining". The preferred reporting items for systematic reviews and meta-analyses (PRISMA) statement was adopted to select and compared distributed learning literature [16]. The PRISMA flow diagram and checklist are slightly modified and presented in the Supplementary Materials (Section A and B) respectively. The last search for distributed machines learning articles was performed on 28-02-2019.

Table 3.1: Examples of machine learning algorithms

	Algorithm	Description	Distributed algorithm available?
Supervised learning	support vector machine (SVM)	An algorithm performing classification tasks by composing hyperplanes in a multidimensional space that separates cases of different class labels. SVM supports both regression and classification tasks and can handle multiple continuous and categorical variables [2,3].	Yes, [2]
	Logistic regression	An algorithm used for discrete values estimation tasks based on a given set of independent variables. In simple words, it predicts the probability of occurrence of an event by fitting data to a logit function that maps probabilities with values lying between 0 and 1. Hence, it is also known as logit regression [78,79].	Yes, [80–82]
	Decision tree	A nonparametric method used for both classification and regression problems. As the name suggests, it uses a tree like decision model by splitting a dataset into two or more subsets based on conditional control statements. It can also be used to visually represent decision making processes[10].	Yes, [83,84]
	Random forest	Random forest is a method performing classification and regression tasks. A random forest is a composition of two initials; forest because it represents a collection of decision trees. Random, because the forest randomly selects observations and features from which it puts up various decision trees, the results are then averaged. Each decision tree in the forest has access to a random set of the training data, choses a class and the most selected class is then the predicted class [10].	No
	KNN	K-nearest neighbors can be used for regression problems; however, it is widely used for classification problems. In KNN the assumption is that similar data elements are close to each other. Given K (positive integer) and a test observation, KNN first groups the K closest elements to the test observation, then in case of regression, return the mean of the K labels or the mode of the K labels in case of classification [10].	Yes, [85,86]

	Algorithm	Description	Distributed algorithm available?
Unsupervised learning	K-means	An algorithm mainly used for clustering in data mining. K-means nomination comes from its functionality; which is partitioning of N observations to K clusters, where each and every observation is part of the cluster with the nearest mean [87].	Yes, [88–90]
	Apriori algorithm	A classical algorithm in data mining. Used for mining frequent item groups and relevant association rules in a transactional database [91].	Yes, [92]
Reinforcement learning	Markov decision processes (MDP)	Introduced in 1950s [93], MDP is a discrete stochastic control process providing a framework for modeling decision making when final outcomes are ambiguous. Given S_o (current state) and S (new state); the decision process is made by steps, the process has a state at each step and the decision maker can choose any available action in S_o. The process then moves randomly into S the new state. The chosen action influences the probability of moving from S_o to S. In other words the next state depends only on the current state not the previous states and the action taken by the decision maker; satisfying the Markovian property, from where comes algorithm's name [94].	No
	Q-learning	Useful for optimization of action selection of any finite MDP. Q-learning algorithm provides agents in a process the ability to know what action to take under what situation [95].	Yes, [96]

3

Table 3.2: Summary of methods and results of distributed machine learning studies grouping more than one healthcare centers

Author / ref	Data and target	Methods & distributed learning approach	Tools	Accomplishments / results
Arthur Jochems [59]	Clinical data from 287 lung cancer patients, treated with curative intent with chemoradiation (CRT) or radiotherapy (RT) alone were collected and stored in five different medical institutes: MAASTRO: (The Netherlands), Jessa (Belgium), Liège (Belgium), Aachen (Germany) and Eindhoven (The Netherlands). Target: predict dyspnea.	A Bayesian network model is adapted for distributed learning using data from five hospitals. Where patient data were extracted and stored in locally in the hospitals. Only the weights are then sent to the master server.	Varian learning portal	AUC = 0.61
Timo M. Deist [21]	Clinical data from 268 NSCLC patients from five different medical institutes: Aachen (Germany), Eindhoven (The Netherlands), Hasselt (Belgium), Liège (Belgium) and Maastricht (The Netherlands). Target: predict dyspnea grade >= 2.	Alternating Direction Method of Multipliers was used to learn SVM models. The data was processed simultaneously in local databases. Then, the updated model parameters are sent to the master machine to compare and update them; check if the learning process has converged. The process is repeated until convergence criteria are met.	Varian Learning Portal	AUC = 0.62 for training set and 0.66 for validation set.
Petr Dluhos [63]	258 first episode schizophrenia patients and 222 healthy controls originating from four datasets were collected: two datasets from University Hospital Brno (Czech Republic), University Medical Center Utrecht (The Netherlands) and the last dataset originates from the Prague Psychiatric Center and Psychiatric hospital Bohnice. Target: classification of patients with First-episode schizophrenia.	All images were preprocessed: normalized, segmented and standardized. Create four local SVM models. Then create multi sample models (joint model and meta model) based on the individual models created previously. This process was repeated four times, by setting each time three training datasets and the remaining one is the validation set.	VBM8 toolbox MATLAB statistics and machine learning toolbox	Joint and meta models had similar classification performance, whereas it was better than performance of local models.

Author / ref	Data and target	Methods & distributed learning approach	Tools	Accomplishments / results
Arthur Jochems [60]	Clinical data from 698 lung cancer patients, treated with curative intent with chemoradiation (CRT) or radiotherapy (RT) alone were collected and stored in two medical institutes: (MAASTRO (The Netherlands) and Michigan University (United States). Target: prediction of NSCLC two-year survival after radiation therapy.	Distributed learning for a Bayesian network using data from three hospitals. The model used the T category and, N category, age, total tumor dose and world health organization performance for predictions.	Varian learning portal	AUC = 0.662 The discriminative performance of centralized and distributed models on the validation set was similar.
Theodora S. Brisimi [62]	Electronic health records from Boston Medical Center of patients with at least one heart-related diagnosis between 2005 and 2010. The data is distributed between 10 hospitals. Target: prediction of heart cardiac events.	Soft-margin l1-regularized sparse SVM classifier.	Not provided	AUC = 0.56. Developed an iterative cluster Primal Dual Splitting (cPDS) algorithm for solving the large-scale sSVM problem in a decentralized fashion. The system then predicts patient's hospitalization for cardiac events in upcoming calendar year. cPDS converges faster than centralized methods.
Luca Tagliaferri [61]	227 variables extracted from thyroid cancer data from six Italian cancer centers, each has 4 properties: name, form, type of field, and levels. Target: prediction of survival and toxicity.	Inferential regression analysis. Learning Analyzer Proxy (module of BOA only in distributed mode) sends algorithms directly to Local Research Proxies, taking back from them only the results of each iteration step, with no need to work with shared data in the Cloud anymore.	COBRA framework	Thyroid COBRA: based on COBRA-Storage System (C-SS). A new software BOA "Beyond Ontology" supporting two different models: Cloud based large database model and Distributed learning model.

3

Search results

A total of 127 articles were identified in PubMed using the search query: ("Distributed Learning" OR "Distributed Machine Learning" OR "privacy preserving data mining"). Six papers were screened, a brief summary of each article is presented in the Table 3.2.

Distributed learning
Distributed learning ensures data safety by only sharing mathematical parameters (or metadata) and not the actual data or in any instance data that might enable tracking back the patient information (such as patient ID, name, date of birth …). In other words, distributed algorithms iteratively analyse separate databases and return the same solution as if data were centralized: essentially sharing research questions and answers between databases instead of data (watch the animation: https://youtu.be/nQpqMIuHyOk). Also, before processing with the learning process researchers must make sure all data have been successfully anonymized and secured by means of hashing algorithms and semantic web techniques respectively as can be seen in Figure 3.2. In addition to postprocessing methods to address the multi-center variabilities [17].

- **Distributed machine learning**

A large quantity of training data is required for machine learning to be applied, especially in outcome modelling where multiple factors influence learning. Provided there is sufficient and appropriate data, machine learning typically results in accurate and generalizable models [18,19]. However, the sensitivity of the personal data greatly hinders the conventional centralized approach to machine learning, whereby all data are gathered in a single data store. Distributed machine learning resolves legal and ethical privacy concerns by learning without the personal data ever leaving the firewall of the medical centers [20].

The euroCAT [21] and ukCAT [22] projects are a proof of distributed learning being successfully implemented into clinical settings to overcome data access restrictions. The purpose of euroCAT project was to predict patient outcomes (e.g., post radiotherapy dyspnea for lung cancer patients) by learning from data stored within clinics without sharing any of the medical data.

- **Distributed deep learning**

Training a deep learning model typically requires thousands to millions of data points and is therefore computationally expensive as well as time consuming. These challenges can be mitigated with different approaches. First, as it is possible to train deep learning models in a parallelized fashion [23], using dedicated hardware (graphics processing units, tensor processing units) [24] reduces the computational time. Second, as the memory of this dedicated hardware is often limited, it is possible to divide the
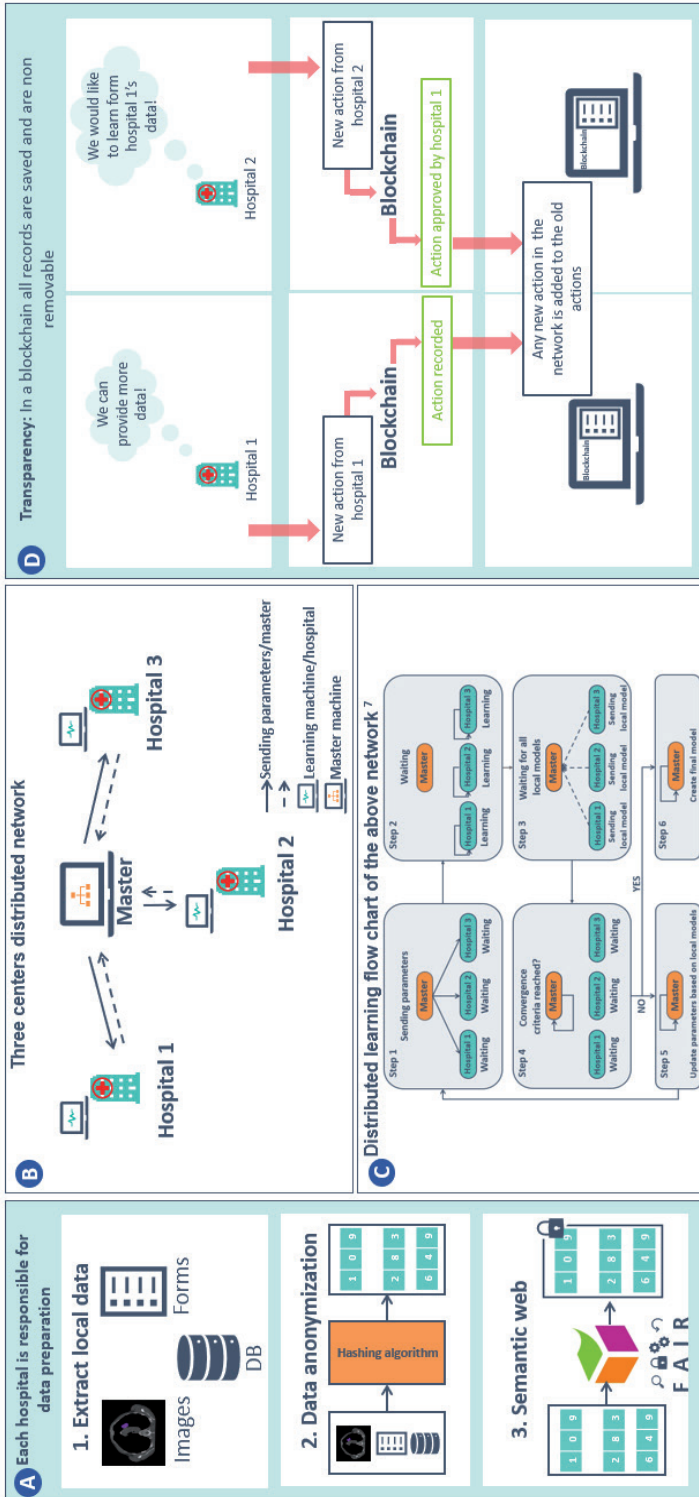
Figure 3.2: Schematic representation of the processes in a transparent distributed learning network. (A) represents the distributed learning network which is composed of three hospitals, each of which is equipped with a learning machine, that can communicate with a master machine responsible for sending model parameters and checking convergence criteria; **(C)** represents a flow chart of the distributed learning network described in (B); **(D)** represents an example of an action that can be tracked by blockchain (designed and implemented according to needs agreed between network members) and keep all network participants aware of any new activity taken in the network.

3

training data into subsets called batches. In this situation, the training process iterates over the batches, only considering the data of one batch at each iteration [25]. On the top of easing the computing burden, using small batches during training improves the model's ability to generalize [26].

These approaches address computation challenges, but do not necessarily preserve data privacy. As for machine learning, deep learning can be distributed to protect patient data [27,28]. Moreover, distributed deep learning also improves computing performance as in the case of wireless sensor networks, where centralized learning is inefficient, both in terms of communication and energy [29,30].

An example of distributed deep learning in the medical domain is that of Chang et al. [31] who deployed a deep learning model across four medical institutions for image classification purposes using three distinct datasets: retinal fundus, mammography, and ImageNet. The results were compared with the same deep learning model trained on centrally hosted data. The comparison showed that the distributed model accuracy is similar to the centrally hosted model [31]. In a different study, McClure et al, developed a distributed deep neural network model, to reproduce Freesurfer[1] brain segmentation. The results demonstrated a performance improvement on the test datasets [32]. Similar to the previous study, a brain tumor segmentation was successfully performed using distributed deep learning across ten institutions (BraTS distribution) [33].

In the matter of distributed deep learning, the training weights are combined to train a final model, and the raw data is never exposed [32,34]. In the case of sharing the locale gradients [23], it might be possible to retrieve estimations of the original data from these gradients. Training the local models on batches may prevent retrieving all the data from the gradients, as these gradients correspond to single batches not all the local data [35]. However, setting an optimal batch size needs to be considered [23], to assure data safety and model's ability to generalize [26,36,37].

- **Privacy and integration of distributed learning networks**

Privacy in a distributed leaning network addresses three main areas; data privacy, the implemented model's privacy, and the model's output privacy. Data privacy is achieved by means of data anonymization and data never leaving the medical institutions. The distributed learning model can be secured by applying differential privacy techniques [38] preventing leakage of weights during the training, and cryptographic techniques [39]. These cryptographic techniques provide a set of multi-party protocols that ensure security of the computations and communication. Once the model is ready, not only can the network participants use it to learn from their data, but this learning should be able to be performed locally and under highly private and secure conditions to protect model's output [21].

---

1   An open-source tool for preprocessing and analyzing (segmentation, thickness estimation …) of human brain MRI images (http://freesurfer.net/fswiki/FreeSurferWiki).

The users of a machine/deep learning model are not necessarily the model's developers. Hence, documentation and the integration of automated data eligibility tests have two important assets:

- The documentation ensures providing a clear view of what the model is designed for, a technical description of the model, and its use.

- The eligibility tests are important to ensure that correct input data is extracted and provided prior to executing the model: In euroCAT [21], a distributed leaning expert installed quality control via data extraction pipelines at every participant point in the network. The pipeline automatically allowed data records fulfilling the model training eligibility criteria to be used in the training. The experts also test the extraction pipeline thoroughly in addition to the machine learning testing. However, there were post processing compensation methods to correct for the variations caused by using different local protocols [17].

Discussion

If one examines oncology, for instance, cancer is evidently one of the greatest challenges facing healthcare. More than 16 million new cancer cases were reported in 2017 alone [40]. This number climbed to 18.1 million cases in 2018 [41]. This increasing number of cancer incidences [42] means that there is undoubtedly sufficient data worldwide to put machine/deep learning to meaningful work. However, as highlighted earlier this requires access to the data and, as also highlighted earlier, distributed learning enables this in a manner that resolves legal and ethical concerns. Nonetheless, integration of distributed learning into healthcare is much slower in comparison to other fields, which raises the question of why this should be? Here, we summarize a set of methodologies to fasten the adoption of distributed learning and provide future directions.

Current state of medical data storage and preprocessing

- **Information communication technology (ICT)**

Every hospital has its own storage devices and architecture [35,36]. In this case, the ICT preparation for distributed learning requires significant energy, time, and manpower, which can be costly. This same process (data acquisition and preprocessing) needs to be repeated for each participating hospital [45] and subsequently development and adoption of medical data standardization protocols need to be developed for this implementation process.

- **Make the data readable: FAIR data principles**

One way to enable a virtuous circle network effect is to embrace another community engaged in synergistic activities (joining a distributed learning network is worthwhile if it links to another large network). The FAIR guiding principles for data management and stewardship (FAIR = findable, accessible, interoperable, re-usable) have gained substantial interest, but delivering scientific protocols and workflows that are aligned with these principles is very significant [46]. A description of FAIR principles is represented in Figure 3.3. Technological solutions are urgently needed that will enable researchers to explore, consume, and produce FAIR data in a reliable and efficient manner, to publish and reuse computational workflows, and to define and share scientific protocols as workflow templates [47]. Such solutions will address emerging concerns about the non-reproducibility of scientific research, particularly in data science (e.g., poorly published data, incomplete workflow descriptions, limited ability to perform meta-analyses, and an overall lack of reproducibility) [48,49]. As workflows are fundamental to research activities, FAIR has broad applicability which is vital in the context of distributed learning with medical data.
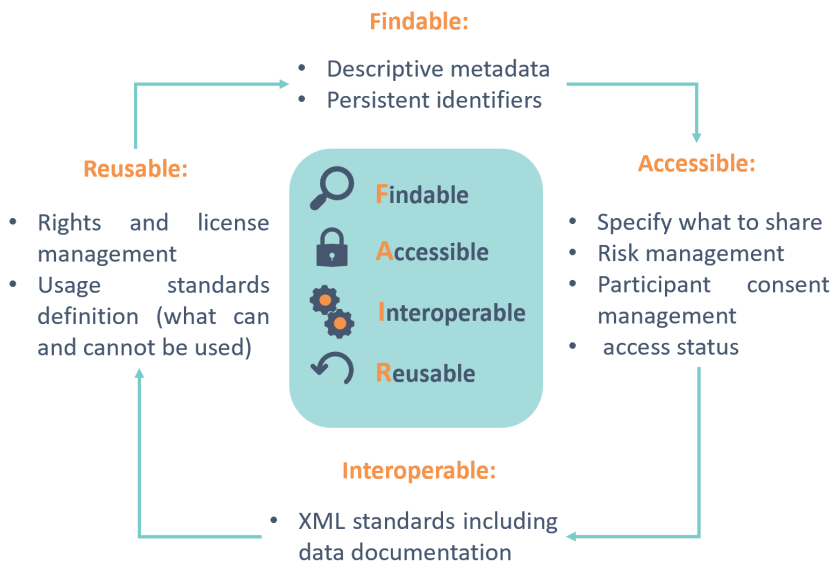

Findable:
- Descriptive metadata
- Persistent identifiers

Reusable:
- Rights and license management
- Usage standards definition (what can and cannot be used)

Findable
Accessible
Interoperable
Reusable

Accessible:
- Specify what to share
- Risk management
- Participant consent management
- access status

Interoperable:
- XML standards including data documentation

Figure 3.3: Description of FAIR principles.

Why not publicly share medical data?

Some studies were conducted trying to facilitate and secure data sharing procedures to encourage related researchers and organizations to publicly share their data and embrace transparency [50], by proposing data sharing procedures and protocols aiming to harmonize regulatory frameworks and research governance [51,52]. Despite the efforts made towards data sharing globalization, the sociocultural issues surround-

ing data sharing remain pertinent [53]. Large clinical trials also face limitations in the data collection capabilities due to limited data storage capacities and manpower. To retrospectively perform additional analysis, all the participating centers need to be contacted again, which is time consuming and delays research [54].

Further, medical institutions prefer not to share patient data to ensure privacy protection [55]. This is, of course, in no small part about ensuring the trust and confidence of patients who display a wide range of sensitivities towards the use of their personal data.

Organizational change management

The adoption of distributed learning will require a change in organizational management (such as making use of newest data standardization techniques and, adapting the roles of employees to more technical oriented tasks as for example data retrieval). Provided knowledge and understanding of proper change management concepts, healthcare providers can implement this latter successfully [56]. Change management principles such as: defining a global vision, networking, and continuous communicating could facilitate the integration of new technologies and bring up the clinical capabilities. However, this process of change management can be complicated because it requires the involvement of multiple healthcare centers from different countries and continents. This diversity can trigger a fear of loss[2], which stems from differences of opinion and regulation [57], and the absence of data standardization -making the processes of data acquisition and preprocessing harder-. Additionally, the lack of knowledge about the new technology which leads to resistance to accept the change and innovation [57,58]. Therefore, it is important to help healthcare organizations understand the need for distributed learning by explaining the context of the change in terms of traditional ways of learning to distributed learning; and a long-term vision of the improvements that it can bring including time and money savings for both hospitals and patients which could in turn improve patient lives. In addition to conducting more studies on research databases to consolidate a proof of safety and quality of distributed models.

As can be seen in Table 3.2, distributed learning has been applied to train different models that can predict different outcomes for a variety of pathologies, including lung cancer [21,59,60], thyroid cancer [61], Heart cardiac events [62], and schizophrenia [63]. This in addition to the continuous development of tools and algorithms facilitating the adoption of distributed learning, such as the variant learning portal, the alternating direction method of multipliers (ADMM) algorithm [2], as well as the application of FAIR data principles. The cited studies provide a proof that distributed learning can ensure patient data privacy and guarantee that accurate models are built that are the equivalent of centralized models.

---

2   One the major factors of financial decision making.

Limitations of the existing distributed learning implementations

A shared limitation of the studies presented in Table 3.2 is that the number of institutes involved in the distributed network is rather small. The size of network varies from four to ten institutions. With few medical institutes involved, the models were trained using the data of few hundred patients only. By promoting the use of distributed learning, it should instead be possible to train the models using data from thousands or even millions of patients.

Future perspectives

An automated monitoring system accessible by the partners or medical centers participating in the distributed learning network can promote transparency, traceability, and trust [64]. Recent advances of information technology such as blockchain can be integrated into a distributed learning network [65]. It allows trusted partners to visualize the history of the transactions and actions taken in the distributed network. This integration of blockchain should help in easing the resistance to the new distributed technology among healthcare workers as it provides both provenance and enforceable governance.

In 2008, Satoshi Nakamoto introduced the concept of peer-to-peer electronic cash system known as Bitcoin [66]. Blockchain was made famous as the public transaction ledger of this cryptocurrency [66,67]. It ensures security by using cryptography in a decentralized, immutable distributed ledger technology [68]. It is easy to manage as it can be made public, whereby any individual can participate, or it can be made private, where all participants are known to each other [69]. It is an efficient monitoring system, as records cannot be deleted from the chain. By these means Blockchain exceeds its application as a cryptocurrency to a permanent trustful tracing system. Figure 3.4 illustrates a visual representation of blockchain.

Boulos et al. demonstrated how blockchain could be used and contribute in healthcare: securing patient information and provider identities, managing health supply chains, monetizing clinical research and data (giving patients the choice to share), processing claims, detecting fraud, and managing prescriptions (replace incorrect and outdated data) [68]. In addition to the abovementioned uses of blockchain, it has been also used to maintain security and scalability of clinical data sharing [70], secure medical records sharing [71], prevent of drug counterfeiting [72] and secure patient's location [73].

It is essential that the use of distributed machine/deep learning and blockchain be harmonized with the available security preserving technologies i.e., continues development and cybersecurity, which begins at the user levels (use strong passwords, connect using only trusted networks …) and ends with more complex IT infrastructures (such as data anonymization, user ID encryption …) [74]. Cybersecurity is a key aspect in preserving
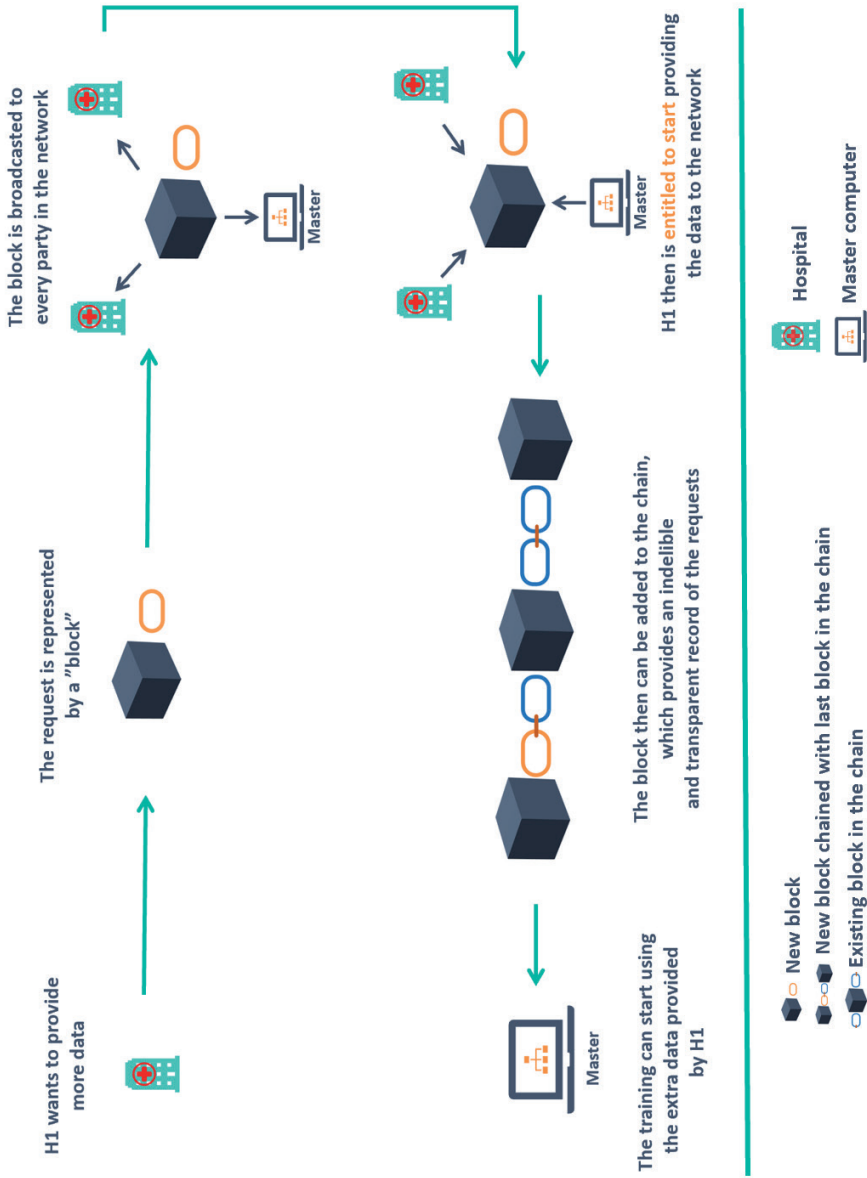
Figure 3.4: Visual representation of blockchain.
https://www.steptoe.com/images/content/1/7/v2/171967/LIT-FebMar18-Feature-Blockchain.pdf.

privacy and, ensuring safety and trust among patients and healthcare systems [75]. The continuous development or post-marketing surveillance can be seen as the set of checks and integrations that should occur when a distributed learning network is launched. This practice should make it possible to identify any weak security measures in the network or non-up-to-date features that may require reimplementation [76,77].

The distributed learning and Blockchain technologies presented here show that there are emerging data science solutions that begin to meet the concerns and shortcomings of the law. The problems of reidentification are greatly reduced and managed through the technologies. Clearly, there are conceptual issues of understanding the impact of these technologies on "privacy", and the relationship between privacy and "confidentiality", but there are significant technical developments for the regulators to consider that could answer a number of their concerns.

Summary

Currently, a combination of regulations and ethics makes it difficult to share data even for scientific research purposes. The issues relate to the legal basis for processing and anonymization. Specifically, there has been reluctance to move away from informed consent as the legal basis for processing towards processing in the public interest, and there are concerns about the reidentification of individuals where data is deidentified and then shared in aggregated environments. A solution could be to allow researchers to train their machine learning programs without the data ever having to leave the clinics which in this paper we have established as distributed learning. This safe practice makes it possible to learn from medical data and can be applied across various medical disciplines. A limitation to its application, however, is that medical centers need to be convinced to participate in such practice, and regulators also need to know suitable safeguards have been established. Moreover, as can be seen in Table 3.2, even with the use of distributed learning, the size of the data learned from remain rather small. In the future the integration of Blockchain technology to distributed learning networks could be considered as it ensures transparency and traceability whilst following FAIR data principles can facilitate the implementation of distributed learning.

Conflict of interest statement

Fadila Zerka and Samir Barakat are salaried employees of Oncoradiomics SA.

Dr Lambin reports grants/sponsored research from Oncoradiomics, ptTheragnostic, advisor (SAB)/presentor fee from Oncoradiomics. Dr Lambin and RL have (minority) shares in the company Oncoradiomics and coinventor of two licensed & approved patents on Radiomics.

3

References

[1]   T.M. Mitchell, Machine learning, International ed., [Reprint.], McGraw-Hill, New York, NY, 1997.

[2]   S. Boyd, Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers, Foundations and Trends® in Machine Learning. 3 (2010) 1–122. https://doi.org/10.1561/2200000016.

[3]   I. Cardoso, E. Almeida, H. Allende-Cid, A. Frery, R. Rangayyan, P. Azevedo-Marques, H. Ramos, Analysis of Machine Learning Algorithms for Diagnosis of Diffuse Lung Diseases, Methods of Information in Medicine. (2018). https://doi.org/10.3414/ME17-02-0023.

[4]   X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, R.M. Summers, ChestX-Ray8: Hospital-Scale Chest X-Ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, Honolulu, HI, 2017: pp. 3462–3471. https://doi.org/10.1109/CVPR.2017.369.

[5]   Y. Ding, J.H. Sohn, M.G. Kawczynski, H. Trivedi, R. Harnish, N.W. Jenkins, D. Lituiev, T.P. Copeland, M.S. Aboian, C. Mari Aparici, S.C. Behr, R.R. Flavell, S.-Y. Huang, K.A. Zalocusky, L. Nardo, Y. Seo, R.A. Hawkins, M. Hernandez Pampaloni, D. Hadley, B.L. Franc, A Deep Learning Model to Predict a Diagnosis of Alzheimer Disease by Using 18 F-FDG PET of the Brain, Radiology. (2018) 180958. https://doi.org/10.1148/radiol.2018180958.

[6]   F. Emmert-Streib, M. Dehmer, A Machine Learning Perspective on Personalized Medicine: An Automized, Comprehensive Knowledge Base with Ontology for Pattern Recognition, Machine Learning and Knowledge Extraction. 1 (2018) 149–156. https://doi.org/10.3390/make1010009.

[7]   T.M. Deist, F.J.W.M. Dankers, G. Valdes, R. Wijsman, I.-C. Hsu, C. Oberije, T. Lustberg, J. van Soest, F. Hoebers, A. Jochems, I. El Naqa, L. Wee, O. Morin, D.R. Raleigh, W. Bots, J.H. Kaanders, J. Belderbos, M. Kwint, T. Solberg, R. Monshouwer, J. Bussink, A. Dekker, P. Lambin, Machine learning algorithms for outcome prediction in (chemo)radiotherapy: An empirical comparison of classifiers, Medical Physics. 45 (2018) 3449–3459. https://doi.org/10.1002/mp.12967.

[8]   P. Lambin, R.G.P.M. van Stiphout, M.H.W. Starmans, E. Rios-Velazquez, G. Nalbantov, H.J.W.L. Aerts, E. Roelofs, W. van Elmpt, P.C. Boutros, P. Granone, V. Valentini, A.C. Begg, D. De Ruysscher, A. Dekker, Predicting outcomes in radiation oncology—multifactorial decision support systems, Nature Reviews Clinical Oncology. 10 (2013) 27–40. https://doi.org/10.1038/nrclinonc.2012.196.

[9]   S. Wang, R.M. Summers, Machine learning and radiology, Medical Image Analysis. 16 (2012) 933–951. https://doi.org/10.1016/j.media.2012.02.005.

[10]   G. James, D. Witten, T. Hastie, R. Tibshirani, An introduction to statistical learning: with applications in R, Corrected at 8th printing, Springer, New York Heidelberg Dordrecht London, 2017.

[11]   R.S. Sutton, A.G. Barto, Reinforcement Learning: An Introduction, (n.d.) 352.

[12]   L. Deng, Deep Learning: Methods and Applications, Foundations and Trends® in Signal Processing. 7 (2014) 197–387. https://doi.org/10.1561/2000000039.

[13]   Y. LeCun, Y. Bengio, G. Hinton, Deep learning, Nature. 521 (2015) 436–444. https://doi.org/10.1038/nature14539.

[14]   C. Garling, Andrew Ng: Why 'Deep Learning' Is a Mandate for Humans, Not Just Machines, Wired. (2015). https://www.wired.com/brandlab/2015/05/andrew-ng-deep-learning-mandate-humans-not-just-machines/ (accessed June 11, 2019).

[15]   F. Pesapane, M. Codari, F. Sardanelli, Artificial intelligence in medical imaging: threat or opportunity? Radiologists again at the forefront of innovation in medicine, Eur Radiol Exp. 2 (2018) 35. https://doi.org/10.1186/s41747-018-0061-6.

[16]   A. Liberati, D.G. Altman, J. Tetzlaff, C. Mulrow, P.C. Gøtzsche, J.P.A. Ioannidis, M. Clarke, P.J. Devereaux, J. Kleijnen, D. Moher, The PRISMA Statement for Reporting Systematic Reviews and Meta-Analyses of Studies That Evaluate Health Care Interventions: Explanation and Elaboration, PLoS Med. 6 (2009) e1000100. https://doi.org/10.1371/journal.pmed.1000100.

[17]   F. Orlhac, F. Frouin, C. Nioche, N. Ayache, I. Buvat, Validation of A Method to Compensate Multicenter Effects Affecting CT Radiomics, Radiology. 291 (2019) 53–59. https://doi.org/10.1148/radiol.2019182023.

[18]   I. Goodfellow, Y. Bengio, A. Courville, Deep Learning, (n.d.). https://www.deeplearningbook.org/ (accessed November 14, 2018).

[19]   P. Lambin, E. Roelofs, B. Reymen, E.R. Velazquez, J. Buijsen, C.M.L. Zegers, S. Carvalho, R.T.H. Leijenaar, G. Nalbantov, C. Oberije, M. Scott Marshall, F. Hoebers, E.G.C. Troost, R.G.P.M. van Stiphout, W. van Elmpt, T. van der Weijden, L. Boersma, V. Valentini, A. Dekker, 'Rapid Learning health care in oncology' – An approach towards decision support systems enabling customised radiotherapy', Radiotherapy and Oncology. 109 (2013) 159–164. https://doi.org/10.1016/j.radonc.2013.07.007.

[20]   T. Lustberg, J. van Soest, A. Jochems, T. Deist, Y. van Wijk, S. Walsh, P. Lambin, A. Dekker, Big Data in radiation therapy: challenges and opportunities, The British Journal of Radiology. 90 (2017) 20160689. https://doi.org/10.1259/bjr.20160689.

[21]   T.M. Deist, A. Jochems, J. van Soest, G. Nalbantov, C. Oberije, S. Walsh, M. Eble, P. Bulens, P. Coucke, W. Dries, A. Dekker, P. Lambin, Infrastructure and distributed learning methodology for privacy-preserving multi-centric rapid learning health care: euroCAT, Clinical and Translational Radiation Oncology. 4 (2017) 24–31. https://doi.org/10.1016/j.ctro.2016.12.004.

[22]   G. Price, M. van Herk, C. Faivre-Finn, Data Mining in Oncology: The ukCAT Project and the Practicalities of Working with Routine Patient Data, Clinical Oncology. 29 (2017) 814–817. https://doi.org/10.1016/j.clon.2017.07.011.

[23]   J. Dean, G. Corrado, R. Monga, K. Chen, M. Devin, M. Mao, M. Ranzato, A. Senior, P. Tucker, K. Yang, Q.V. Le, A.Y. Ng, Large Scale Distributed Deep Networks, (n.d.) 9.

[24]   D. Cireşan, U. Meier, J. Schmidhuber, Multi-column Deep Neural Networks for Image Classification, ArXiv:1202.2745 [Cs]. (2012). http://arxiv.org/abs/1202.2745 (accessed February 28, 2019).

[25]   P.M. Radiuk, Impact of Training Set Batch Size on the Performance of Convolutional Neural Networks for Diverse Datasets, Information Technology and Management Science. 20 (2017). https://doi.org/10.1515/itms-2017-0003.

[26]   N.S. Keskar, D. Mudigere, J. Nocedal, M. Smelyanskiy, P.T.P. Tang, On Large-Batch Training for Deep Learning: Generalization Gap and Sharp Minima, ArXiv:1609.04836 [Cs, Math]. (2016). http://arxiv.org/abs/1609.04836 (accessed June 12, 2019).

[27]   N. Papernot, M. Abadi, Ú. Erlingsson, I. Goodfellow, K. Talwar, Semi-supervised Knowledge Transfer for Deep Learning from Private Training Data, ArXiv:1610.05755 [Cs, Stat]. (2016). http://arxiv.org/abs/1610.05755 (accessed June 6, 2019).

[28]   R. Shokri, V. Shmatikov, Privacy-Preserving Deep Learning, in: Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security - CCS '15, ACM Press, Denver, Colorado, USA, 2015: pp. 1310–1321. https://doi.org/10.1145/2810103.2813687.

[29]   J.B. Predd, S.B. Kulkarni, H.V. Poor, Distributed learning in wireless sensor networks, IEEE Signal Processing Magazine. 23 (2006) 56–69. https://doi.org/10.1109/MSP.2006.1657817.

[30]   X. Ji, C. Hou, Y. Hou, F. Gao, S. Wang, A Distributed Learning Method for ℓ1-Regularized Kernel Machine over Wireless Sensor Networks, Sensors. 16 (2016) 1021. https://doi.org/10.3390/s16071021.

[31]   K. Chang, N. Balachandar, C. Lam, D. Yi, J. Brown, A. Beers, B. Rosen, D.L. Rubin, J. Kalpathy-Cramer, Distributed deep learning networks among institutions for medical imaging, Journal of the American Medical Informatics Association. 25 (2018) 945–954. https://doi.org/10.1093/jamia/ocy017.

[32]   P. McClure, C.Y. Zheng, J. Kaczmarzyk, J. Rogers-Lee, S. Ghosh, D. Nielson, P.A. Bandettini, F. Pereira, Distributed Weight Consolidation: A Brain Segmentation Case Study, (n.d.) 11.

[33]   M.J. Sheller, G.A. Reina, B. Edwards, J. Martin, S. Bakas, Multi-Institutional Deep Learning Modeling Without Sharing Patient Data: A Feasibility Study on Brain Tumor Segmentation, ArXiv:1810.04304 [Cs, Stat]. (2018). http://arxiv.org/abs/1810.04304 (accessed January 7, 2019).

[34]   W. Li, F. Milletarì, D. Xu, N. Rieke, J. Hancox, W. Zhu, M. Baust, Y. Cheng, S. Ourselin, M.J. Cardoso, A. Feng, Privacy-preserving Federated Brain Tumour Segmentation, ArXiv:1910.00962 [Cs]. (2019). http://arxiv.org/abs/1910.00962 (accessed October 18, 2019).

[35]   M. Abadi, A. Chu, I. Goodfellow, H.B. McMahan, I. Mironov, K. Talwar, L. Zhang, Deep Learning with Differential Privacy, Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security - CCS'16. (2016) 308–318. https://doi.org/10.1145/2976749.2978318.

[36]   D. Mishkin, N. Sergievskiy, J. Matas, Systematic evaluation of CNN advances on the ImageNet, Computer Vision and Image Understanding. 161 (2017) 11–19. https://doi.org/10.1016/j.cviu.2017.05.007.

[37]   T. Lin, S.U. Stich, K.K. Patel, M. Jaggi, Don't Use Large Mini-Batches, Use Local SGD, ArXiv:1808.07217 [Cs, Stat]. (2018). http://arxiv.org/abs/1808.07217 (accessed June 14, 2019).

[38]   A. Biryukov, C. De Cannière, W.E. Winkler, C.C. Aggarwal, M. Kuhn, L. Bouganim, Y. Guo, B. Preneel, G. Bleumer, T. Helleseth, R. Canetti, M. Varia, C. Peters, B. Kaliski, Y. Desmedt, G. Kesidis, M. De Soete, G. Bleumer, B. Schoenmakers, A. Biryukov, C. Adams, A. Biryukov, E. Biham, T. Caddy, C. Dwork, A. Biryukov, M. Just, M. Finiasz, N. Sendrier, K. Sako, P. Gaborit, N. Sendrier, D. Boneh, C. Cachin, D. Gordon, N. Li, S.D.C. di Vimercati, M. Videau, S.N. Yanushkevich, D.V. Popel, B.B.H. Kang, J. Ioannidis, M. Christodorescu, V. Ganapathy, B.B.H. Kang, A. Srivastava, S.W. Smith, Discretionary Access Control Policies (DAC), in: H.C.A. van Tilborg, S. Jajodia (Eds.), Encyclopedia of Cryptography and Security, Springer US, Boston, MA, 2011: pp. 356–358. https://doi.org/10.1007/978-1-4419-5906-5_817.

[39]   B. Pinkas, Cryptographic techniques for privacy-preserving data mining, SIGKDD Explor. Newsl. 4 (2002) 12–19. https://doi.org/10.1145/772862.772865.

[40]   R.L. Siegel, K.D. Miller, A. Jemal, Cancer statistics, 2017, CA: A Cancer Journal for Clinicians. 67 (2017) 7–30. https://doi.org/10.3322/caac.21387.

[41]   F. Bray, J. Ferlay, I. Soerjomataram, R.L. Siegel, L.A. Torre, A. Jemal, Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries, CA: A Cancer Journal for Clinicians. 68 (2018) 394–424. https://doi.org/10.3322/caac.21492.

[42]   R. Siegel, C. DeSantis, K. Virgo, K. Stein, A. Mariotto, T. Smith, D. Cooper, T. Gansler, C. Lerro, S. Fedewa, C. Lin, C. Leach, R.S. Cannady, H. Cho, S. Scoppa, M. Hachey, R. Kirch, A. Jemal, E. Ward, Cancer treatment and survivorship statistics, 2012, CA: A Cancer Journal for Clinicians. 62 (2012) 220–241. https://doi.org/10.3322/caac.21149.

[43]   E.H. Shortliffe, G.O. Barnett, Medical Data: Their Acquisition, Storage, and Use, in: E.H. Shortliffe, L.E. Perreault (Eds.), Medical Informatics, Springer New York, New York, NY, 2001: pp. 41–75. https://doi.org/10.1007/978-0-387-21721-5_2.

[44]   M. Shabani, D. Vears, P. Borry, Raw Genomic Data: Storage, Access, and Sharing, Trends in Genetics. 34 (2018) 8–10. https://doi.org/10.1016/j.tig.2017.10.004.

[45]   S.G. Langer, Challenges for Data Storage in Medical Imaging Research, Journal of Digital Imaging. 24 (2011) 203–207. https://doi.org/10.1007/s10278-010-9311-8.

[46]   M.D. Wilkinson, M. Dumontier, Ij.J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.-W. Boiten, L.B. da Silva Santos, P.E. Bourne, J. Bouwman, A.J. Brookes, T. Clark, M. Crosas, I. Dillo, O. Dumon, S. Edmunds, C.T. Evelo, R. Finkers, A. Gonzalez-Beltran, A.J.G. Gray, P. Groth, C. Goble, J.S. Grethe, J. Heringa, P.A.C. 't Hoen, R. Hooft, T. Kuhn, R. Kok, J. Kok, S.J. Lusher, M.E. Martone, A. Mons, A.L. Packer, B. Persson, P. Rocca-Serra, M. Roos, R. van Schaik, S.-A. Sansone, E. Schultes, T. Sengstag, T. Slater, G. Strawn, M.A. Swertz, M. Thompson, J. van der Lei, E. van Mulligen, J. Velterop, A. Waagmeester, P. Wittenburg, K. Wolstencroft, J. Zhao, B. Mons, The FAIR Guiding Principles for scientific data management and stewardship, Scientific Data. 3 (2016) 160018. https://doi.org/10.1038/sdata.2016.18.
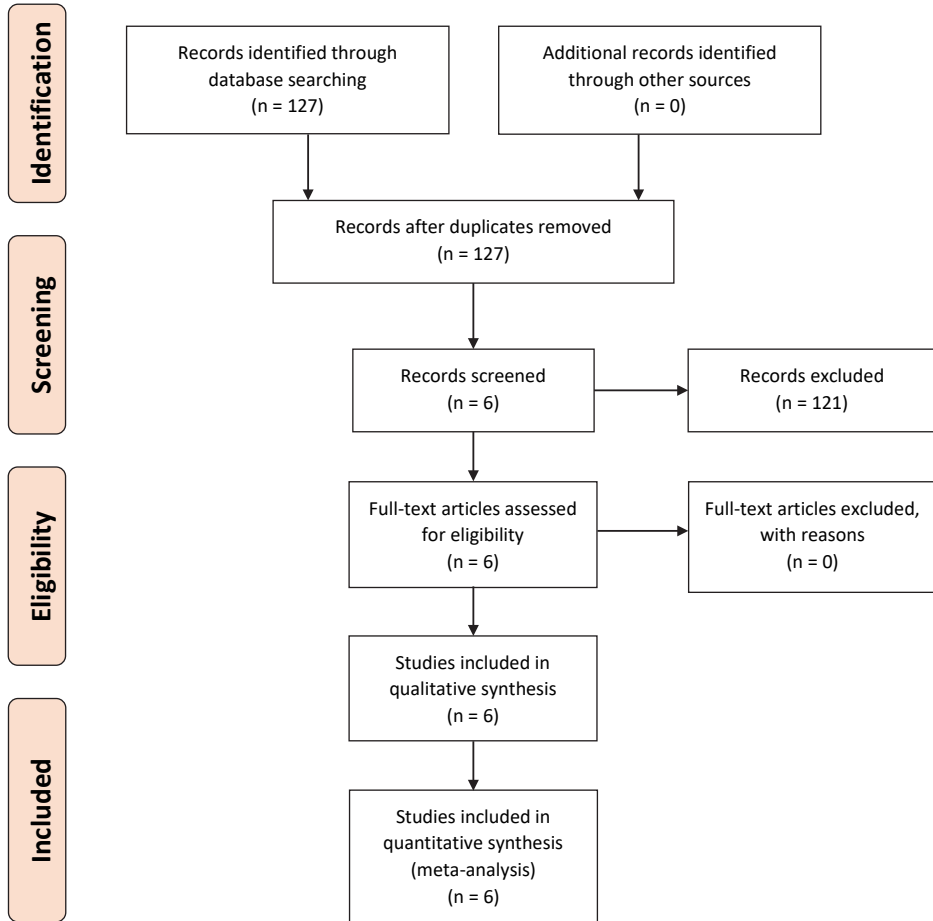
3

[47] M.D. Wilkinson, S.-A. Sansone, E. Schultes, P. Doorn, L.O. Bonino da Silva Santos, M. Dumontier, A design framework and exemplar metrics for FAIRness, Scientific Data. 5 (2018) 180118. https://doi. org/10.1038/sdata.2018.118.

[48] M. Dumontier, A.J.G. Gray, M.S. Marshall, V. Alexiev, P. Ansell, G. Bader, J. Baran, J.T. Bolleman, A. Callahan, J. Cruz-Toledo, P. Gaudet, E.A. Gombocz, A.N. Gonzalez-Beltran, P. Groth, M. Haendel, M. Ito, S. Jupp, N. Juty, T. Katayama, N. Kobayashi, K. Krishnaswami, C. Laibe, N. Le Novère, S. Lin, J. Malone, M. Miller, C.J. Mungall, L. Rietveld, S.M. Wimalaratne, A. Yamaguchi, The health care and life sciences community profile for dataset descriptions, PeerJ. 4 (2016) e2331. https://doi.org/10.7717/peerj.2331.

[49] K.M. Jagodnik, S. Koplev, S.L. Jenkins, L. Ohno-Machado, B. Paten, S.C. Schurer, M. Dumontier, R. Verborgh, A. Bui, P. Ping, N.J. McKenna, R. Madduri, A. Pillai, A. Ma'ayan, Developing a framework for digital objects in the Big Data to Knowledge (BD2K) commons: Report from the Commons Framework Pilots workshop, Journal of Biomedical Informatics. 71 (2017) 49–57. https://doi. org/10.1016/j.jbi.2017.05.006.

[50] J.R. Polanin, M. Terzian, A Data-Sharing Agreement Helps to Increase Researchers' Willingness to Share Primary Data: Results from a Randomized Controlled Trial, Journal of Clinical Epidemiology. 0 (2018). https://doi.org/10.1016/j.jclinepi.2018.10.006.

[51] D.R. Azzariti, E.R. Riggs, A. Niehaus, L.L. Rodriguez, E.M. Ramos, B. Kattman, M.J. Landrum, C.L. Martin, H.L. Rehm, Points to consider for sharing variant-level information from clinical genetic testing with ClinVar, Cold Spring Harb Mol Case Stud. 4 (2018) a002345. https://doi.org/10.1101/mcs.a002345.

[52] S. Boué, M. Byrne, A.W. Hayes, J. Hoeng, M.C. Peitsch, Embracing Transparency Through Data Sharing, International Journal of Toxicology. (2018) 109158181880388. https://doi. org/10.1177/1091581818803880.

[53] J.-B. Poline, J.L. Breeze, S. Ghosh, K. Gorgolewski, Y.O. Halchenko, M. Hanke, C. Haselgrove, K.G. Helmer, D.B. Keator, D.S. Marcus, R.A. Poldrack, Y. Schwartz, J. Ashburner, D.N. Kennedy, Data sharing in neuroimaging research, Frontiers in Neuroinformatics. 6 (2012). https://doi.org/10.3389/fninf.2012.00009.

[54] F.T. Cutts, G. Enwere, S.M.A. Zaman, F.G. Yallop, Operational Challenges in Large Clinical Trials: Examples and Lessons Learned from the Gambia Pneumococcal Vaccine Trial, PLoS Clin Trials. 1 (2006). https://doi.org/10.1371/journal.pctr.0010016.

[55] W. Xia, Z. Wan, Z. Yin, J. Gaupp, Y. Liu, E.W. Clayton, M. Kantarcioglu, Y. Vorobeychik, B.A. Malin, It's all in the timing: calibrating temporal penalties for biomedical data sharing, Journal of the American Medical Informatics Association. 25 (2018) 25–31. https://doi.org/10.1093/jamia/ocx101.

[56] H. Fleishon, L.R. Muroff, S.S. Patel, Change Management for Radiologists, Journal of the American College of Radiology. 14 (2017) 1229–1233. https://doi.org/10.1016/j.jacr.2017.02.053.

[57] R. Delaney, R. D'Agostino, The Challenges of Integrating New Technology into an Organization, (n.d.) 38.

[58] A. Agboola, R. Salawu, Managing Deviant Behavior and Resistance to Change, International Journal of Business and Management. 6 (2010) p235. https://doi.org/10.5539/ijbm.v6n1p235.

[59] A. Jochems, T.M. Deist, J. van Soest, M. Eble, P. Bulens, P. Coucke, W. Dries, P. Lambin, A. Dekker, Distributed learning: Developing a predictive model based on data from multiple hospitals without data leaving the hospital – A real life proof of concept, Radiotherapy and Oncology. 121 (2016) 459–467. https://doi.org/10.1016/j.radonc.2016.10.002.

[60] A. Jochems, T.M. Deist, I. El Naqa, M. Kessler, C. Mayo, J. Reeves, S. Jolly, M. Matuszak, R. Ten Haken, J. van Soest, C. Oberije, G. Faivre-Finn, G. Price, D. de Ruysscher, P. Lambin, A. Dekker, Developing and Validating a Survival Prediction Model for NSCLC Patients Through Distributed Learning Across 3 Countries, International Journal of Radiation Oncology*Biology*Physics. 99 (2017) 344–352. https://doi.org/10.1016/j.ijrobp.2017.04.021.

[61] L. Tagliaferri, C. Gobitti, G.F. Colloca, L. Boldrini, E. Farina, C. Furlan, F. Paiar, F. Vianello, M. Basso, L. Cerizza, F. Monari, G. Simontacchi, M.A. Gambacorta, J. Lenkowicz, N. Dinapoli, V. Lanzotti, R. Mazzarotto, E. Russi, M. Mangoni, A new standardized data collection system for interdisciplinary thyroid cancer management: Thyroid COBRA, European Journal of Internal Medicine. 53 (2018) 73–78. https://doi.org/10.1016/j.ejim.2018.02.012.

[62] T.S. Brisimi, R. Chen, T. Mela, A. Olshevsky, I.Ch. Paschalidis, W. Shi, Federated learning of predictive models from federated Electronic Health Records, International Journal of Medical Informatics. 112 (2018) 59–67. https://doi.org/10.1016/j.ijmedinf.2018.01.007.

[63]   P. Dluhoš, D. Schwarz, W. Cahn, N. van Haren, R. Kahn, F. Španiel, J. Horáček, T. Kašpárek, H. Schnack, Multi-center machine learning in imaging psychiatry: A meta-model approach, NeuroImage. 155 (2017) 10–24. https://doi.org/10.1016/j.neuroimage.2017.03.027.

[64]   V. Dhillon, D. Metcalf, M. Hooper, Blockchain in Health Care, in: V. Dhillon, D. Metcalf, M. Hooper (Eds.), Blockchain Enabled Applications: Understand the Blockchain Ecosystem and How to Make It Work for You, Apress, Berkeley, CA, 2017: pp. 125–138. https://doi.org/10.1007/978-1-4842-3081-7_9.

[65]   S. Lugan, P. Desbordes, L.X.R. Tormo, A. Legay, B. Macq, Secure Architectures Implementing Trusted Coalitions for Blockchained Distributed Learning (TCLearn), ArXiv:1906.07690 [Cs, Stat]. (2019). http://arxiv.org/abs/1906.07690 (accessed August 30, 2019).

[66]   S. Nakamoto, Bitcoin: A Peer-to-Peer Electronic Cash System, (n.d.) 9.

[67]   W.J. Gordon, C. Catalini, Blockchain Technology for Healthcare: Facilitating the Transition to Patient-Driven Interoperability, Computational and Structural Biotechnology Journal. 16 (2018) 224–230. https://doi.org/10.1016/j.csbj.2018.06.003.

[68]   M.N. Kamel Boulos, J.T. Wilson, K.A. Clauson, Geospatial blockchain: promises, challenges, and scenarios in health and healthcare, International Journal of Health Geographics. 17 (2018). https://doi.org/10.1186/s12942-018-0144-x.

[69]   C. Pirtle, J. Ehrenfeld, Blockchain for Healthcare: The Next Generation of Medical Records?, Journal of Medical Systems. 42 (2018). https://doi.org/10.1007/s10916-018-1025-3.

[70]   P. Zhang, J. White, D.C. Schmidt, G. Lenz, S.T. Rosenbloom, FHIRChain: Applying Blockchain to Securely and Scalably Share Clinical Data, Computational and Structural Biotechnology Journal. 16 (2018) 267–278. https://doi.org/10.1016/j.csbj.2018.07.004.

[71]   A. Dubovitskaya, Z. Xu, S. Ryu, M. Schumacher, F. Wang, Secure and Trustable Electronic Medical Records Sharing using Blockchain, AMIA Annu Symp Proc. 2017 (2017) 650–659.

[72]   S. Vruddhula, Application of on-dose identification and blockchain to prevent drug counterfeiting, Pathogens and Global Health. 112 (2018) 161–161. https://doi.org/10.1080/20477724.2018.1503268.

[73]   Y. Ji, J. Zhang, J. Ma, C. Yang, X. Yao, BMPLS: Blockchain-Based Multi-level Privacy-Preserving Location Sharing Scheme for Telecare Medical Information Systems, Journal of Medical Systems. 42 (2018). https://doi.org/10.1007/s10916-018-0998-2.

[74]   L. Coventry, D. Branley, Cybersecurity in healthcare: A narrative review of trends, threats and ways forward, Maturitas. 113 (2018) 48–52. https://doi.org/10.1016/j.maturitas.2018.04.008.

[75]   M.S. Jalali, J.P. Kaiser, Cybersecurity in Hospitals: A Systematic, Organizational Perspective, Journal of Medical Internet Research. 20 (2018) e10059. https://doi.org/10.2196/10059.

[76]   V. Vlahović-Palčevski, D. Mentzer, Postmarketing Surveillance, in: H.W. Seyberth, A. Rane, M. Schwab (Eds.), Pediatric Clinical Pharmacology, Springer Berlin Heidelberg, Berlin, Heidelberg, 2011: pp. 339–351. https://doi.org/10.1007/978-3-642-20195-0_17.

[77]   R. Parkash, B. Thibault, F. Philippon, R. Yee, E. Stephenson, J. Healey, A. Krahn, D. Exner, C. Simpson, E. Crystal, P. Nery, V. Essebag, L. Sterns, A. Tang, G. Wells, Canadian Registry of Implantable Electronic Device Outcomes: Surveillance of High-Voltage Leads, Canadian Journal of Cardiology. 34 (2018) 808–811. https://doi.org/10.1016/j.cjca.2018.02.009.

[78]   E.B. Ing, R. Ing, The Use of a Nomogram to Visually Interpret a Logistic Regression Prediction Model for Giant Cell Arteritis, Neuro-Ophthalmology. 42 (2018) 284–286. https://doi.org/10.1080/01658107.2018.1425728.

[79]   M. Tirzīte, M. Bukovskis, G. Strazda, N. Jurka, I. Taivans, Detection of lung cancer with electronic nose and logistic regression analysis, Journal of Breath Research. (2018). https://doi.org/10.1088/1752-7163/aae1b8.

[80]   Z. Ji, X. Jiang, S. Wang, L. Xiong, L. Ohno-Machado, Differentially private distributed logistic regression using private and public data, BMC Medical Genomics. 7 (2014) S14. https://doi.org/10.1186/1755-8794-7-S1-S14.

[81]   W. Jiang, P. Li, S. Wang, Y. Wu, M. Xue, L. Ohno-Machado, X. Jiang, WebGLORE: a Web service for Grid LOgistic REgression, Bioinformatics. 29 (2013) 3238–3240. https://doi.org/10.1093/bioinformatics/btt559.

[82]   wang shuang, jiang Xiaoqian, wu yuan, cui lijuan, cheng Samuel, L. Ohno-Machado, EXpectation Propagation LOgistic REgRession (EXPLORER): Distributed Privacy-Preserving Online Model Learning, (n.d.). https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3676314/ (accessed December 24, 2018).

3

[83]   A. Desai, S. Chaudhary, Distributed Decision Tree, in: Proceedings of the Ninth Annual ACM India Conference on - ACM COMPUTE '16, ACM Press, Gandhinagar, India, 2016: pp. 43–50. https://doi.org/10.1145/2998476.2998478.

[84]   D. Caragea, A. Silvescu, V. Honavar, Decision Tree Induction from Distributed Heterogeneous Autonomous Data Sources, in: A. Abraham, K. Franke, M. Köppen (Eds.), Intelligent Systems Design and Applications, Springer Berlin Heidelberg, 2003: pp. 341–350.

[85]   E. Plaku, L.E. Kavraki, Distributed Computation of the knn Graph for Large High-Dimensional Point Sets, J Parallel Distrib Comput. 67 (2007) 346–359. https://doi.org/10.1016/j.jpdc.2006.10.004.

[86]   L. Xiong, S. Chitti, L. Liu, Mining multiple private databases using a kNN classifier, in: Proceedings of the 2007 ACM Symposium on Applied Computing - SAC '07, ACM Press, Seoul, Korea, 2007: p. 435. https://doi.org/10.1145/1244002.1244102.

[87]   Z. Huang, Extensions to the k-Means Algorithm for Clustering Large Data Sets with Categorical Values, Data Mining and Knowledge Discovery. 2 (1998) 283–304. https://doi.org/10.1023/A:1009769707641.

[88]   G. Jagannathan, R.N. Wright, Privacy-preserving distributed k-means clustering over arbitrarily partitioned data, in: Proceeding of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining - KDD '05, ACM Press, Chicago, Illinois, USA, 2005: p. 593. https://doi.org/10.1145/1081870.1081942.

[89]   R. Jin, A. Goswami, G. Agrawal, Fast and exact out-of-core and distributed k-means clustering, Knowledge and Information Systems. 10 (2006) 17–40. https://doi.org/10.1007/s10115-005-0210-0.

[90]   G. Jagannathan, K. Pillaipakkamnatt, R.N. Wright, A New Privacy-Preserving Distributed k -Clustering Algorithm, in: Proceedings of the 2006 SIAM International Conference on Data Mining, Society for Industrial and Applied Mathematics, 2006: pp. 494–498. https://doi.org/10.1137/1.9781611972764.47.

[91]   Yanbin Ye, Chia-Chu Chiang, A Parallel Apriori Algorithm for Frequent Itemsets Mining, in: Fourth International Conference on Software Engineering Research, Management and Applications (SERA'06), IEEE, Seattle, WA, USA, 2006: pp. 87–94. https://doi.org/10.1109/SERA.2006.6.

[92]   D.W. Cheung, V.T. Ng, A.W. Fu, Yongjian Fu, Efficient mining of association rules in distributed databases, IEEE Transactions on Knowledge and Data Engineering. 8 (1996) 911–922. https://doi.org/10.1109/69.553158.

[93]   R. Bellman, A Markovian Decision Process, Indiana Univ. Math. J. 6 (1957) 679--684.

[94]   M.L. Puterman, Markov Decision Processes: Discrete Stochastic Dynamic Programming, John Wiley & Sons, 2014.

[95]   C.J.C.H. Watkins, P. Dayan, Q-learning, Machine Learning. 8 (1992) 279–292. https://doi.org/10.1007/BF00992698.

[96]   M. Lauer, M. Riedmiller, An Algorithm for Distributed Reinforcement Learning in Cooperative Multi-Agent Systems, in: In Proceedings of the Seventeenth International Conference on Machine Learning, Morgan Kaufmann, 2000: pp. 535–542.

Supplementary materials

Section A
PRISMA 2009 Flow Diagram

3

Section B
PRISMA 2009 Checklist

Section/topic	#	Checklist item	Reported on page #
TITLE			
Title	1	Identify the report as a systematic review, meta-analysis, or both.	1
ABSTRACT			
Structured summary	2	Provide a structured summary including, as applicable: background; objectives; data sources; study eligibility criteria, participants, and interventions; study appraisal and synthesis methods; results; limitations; conclusions and implications of key findings; systematic review registration number.	3
INTRODUCTION			
Rationale	3	Describe the rationale for the review in the context of what is already known.	4-6
Objectives	4	Provide an explicit statement of questions being addressed with reference to participants, interventions, comparisons, outcomes, and study design (PICOS).	7
METHODS			
Protocol and registration	5	Indicate if a review protocol exists, if and where it can be accessed (e.g., Web address), and, if available, provide registration information including registration number.	9
Eligibility criteria	6	Specify study characteristics (e.g., PICOS, length of follow-up) and report characteristics (e.g., years considered, language, publication status) used as criteria for eligibility, giving rationale.	9
Information sources	7	Describe all information sources (e.g., databases with dates of coverage, contact with study authors to identify additional studies) in the search and date last searched.	9
Search	8	Present full electronic search strategy for at least one database, including any limits used, such that it could be repeated.	9
Study selection	9	State the process for selecting studies (i.e., screening, eligibility, included in systematic review, and, if applicable, included in the meta-analysis).	9 (and Supplementary Materials (Section A))
Data collection process	10	Describe method of data extraction from reports (e.g., piloted forms, independently, in duplicate) and any processes for obtaining and confirming data from investigators.	9
Data items	11	List and define all variables for which data were sought (e.g., PICOS, funding sources) and any assumptions and simplifications made.	N/A
Risk of bias in individual studies	12	Describe methods used for assessing risk of bias of individual studies (including specification of whether this was done at the study or outcome level), and how this information is to be used in any data synthesis.	N/A

Section/topic	#	Checklist item	Reported on page #
Summary measures	13	State the principal summary measures (e.g., risk ratio, difference in means).	N/A
Synthesis of results	14	Describe the methods of handling data and combining results of studies, if done, including measures of consistency (e.g., I²) for each meta-analysis.	9
Risk of bias across studies	15	Specify any assessment of risk of bias that may affect the cumulative evidence (e.g., publication bias, selective reporting within studies).	N/A
Additional analyses	16	Describe methods of additional analyses (e.g., sensitivity or subgroup analyses, meta-regression), if done, indicating which were pre-specified.	N/A
RESULTS			
Study selection	17	Give numbers of studies screened, assessed for eligibility, and included in the review, with reasons for exclusions at each stage, ideally with a flow diagram.	9 (and Supplementary Materials (Section A))
Study characteristics	18	For each study, present characteristics for which data were extracted (e.g., study size, PICOS, follow-up period) and provide the citations.	28-32
Risk of bias within studies	19	Present data on risk of bias of each study and, if available, any outcome level assessment (see item 12).	N/A
Results of individual studies	20	For all outcomes considered (benefits or harms), present, for each study: (a) simple summary data for each intervention group (b) effect estimates and confidence intervals, ideally with a forest plot.	28-32
Synthesis of results	21	Present results of each meta-analysis done, including confidence intervals and measures of consistency.	28-32
Risk of bias across studies	22	Present results of any assessment of risk of bias across studies (see Item 15).	N/A
Additional analysis	23	Give results of additional analyses, if done (e.g., sensitivity or subgroup analyses, meta-regression [see Item 16]).	N/A
DISCUSSION			
Summary of evidence	24	Summarize the main findings including the strength of evidence for each main outcome; consider their relevance to key groups (e.g., healthcare providers, users, and policy makers).	14
Limitations	25	Discuss limitations at study and outcome level (e.g., risk of bias), and at review-level (e.g., incomplete retrieval of identified research, reporting bias).	18
Conclusions	26	Provide a general interpretation of the results in the context of other evidence, and implications for future research.	21
FUNDING			
Funding	27	Describe sources of funding for the systematic review and other support (e.g., supply of data); role of funders for the systematic review.	22

3

Sequential learning and Blockchain

4

Blockchain for privacy preserving and trustworthy distributed machine learning in multicentric medical imaging (C-DistriM)

Fadila Zerka | Visara Urovi | Akshayaa Vaidyanathan | Samir Barakat |
Ralph T.H. Leijenaar | Sean Walsh | Hanif Gabrani-Juma |
Benjamin Miraglio | Henry C. Woodruff | Michel Dumontier |
Philippe Lambin

Abstract

The utility of artificial intelligence (AI) in healthcare strongly depends upon the quality of the data used to build models, and the confidence in the predictions they generate. Access to sufficient amounts of high-quality data to build accurate and reliable models remains problematic owing to substantive legal and ethical constraints in making clinically relevant research data available offsite. New technologies such as distributed learning offer a pathway forward, but unfortunately tend to suffer from a lack of transparency, which undermines trust in what data are used for the analysis. To address such issues, we hypothesized that, a novel distributed learning that combines sequential distributed learning with a blockchain-based platform, namely chained distributed machine learning C-DistriM, would be feasible and would give a similar result as a standard centralized approach. C-DistriM enables health centers to dynamically participate in training distributed learning models. We demonstrate C-DistriM using the NSCLC-Radiomics open data to predict two-year lung-cancer survival. A comparison of the performance of this distributed solution, evaluated in six different scenarios, and the centralized approach, showed no statistically significant difference (AUCs between central and distributed models), all DeLong tests yielded p-value > 0.05. This methodology removes the need to blindly trust the computation in one specific server on a distributed learning network. This fusion of blockchain and distributed learning serves as a proof-of-concept to increase transparency, trust, and ultimately accelerate the adoption of AI in multicentric studies. We conclude that our blockchain-based model for sequential training on distributed datasets is a feasible approach, provides equivalent performance to the centralized approach.

Introduction

The application of artificial intelligence (AI) algorithms in medical imaging has evolved from machine learning that is able to learn from quantitative (Radiomics) features to deep learning algorithms, mostly convolutional neural networks (CNN), that are in turn able to learn complex non-linear features from medical imaging and inform about diagnosis, prognosis, and personalize treatment options [1–5]. The CNN algorithms showed a great performance when applied to medical imaging [6,7]. Ultimately the ability to successfully generalize an AI algorithm is influenced by the quality (volume, veracity, variety, and velocity -4Vs) of the training data [8]. As the data quality improves [9], a similar trend is seen in both performance and generalizability. Typically, a solitary medical center does not have sufficient quality data for the specific task at hand to implement high-performance AI for use in other sites. The conventional approach to access high quality data in healthcare is through multicentric studies, however, recent legal and ethical considerations (e.g., general data protection regulation (GDPR) and health insurance portability and accountability act (HIPAA)) have now made multicentric studies with centralized databases problematic [10]. One potential way to address this challenge is to share the training workload of machine learning models rather than centralize the data, originating from multiple institutions. This approach, proposed in 2013, is known as distributed learning (federated learning) [11,12].

Distributed learning – a fusion of machine learning and distributed computing – allows machine learning models to be trained on multiple siloed datasets without the need for patient data to leave the firewalls of each database [13]. Distributed learning preserves privacy by design, by sharing model weights for subsequent training cycles instead of privacy sensitive data. Distributed learning has been successfully applied to train machine learning models using data originating from multiple medical centers [11,14–16] on a global scale, producing models with equivalent performance to centralized data training approach [17].

A distributed learning network involves multiple partners. Within the network, each partner is connected to a central coordinator (i.e., the master server) that initializes and aggregates the learning. This design however is vulnerable to malicious or (un) intentional misuse of the network, as researchers have demonstrated it is possible to retrieve sensitive patient information from the shared weights of the model [18]. Furthermore, it is impossible for each partner to monitor the quality of the data provided by others within the network. In essence, this approach requires collaborators to blindly trust the master server. Given the risks associated with this design, elevating the transparency and traceability of the data and learning may improve usability and confidence of this approach.

4

Blockchain is a technology utilizing cryptographic hashing techniques to maintain a distributed data structure that stores information in an append-only manner. The integration of a blockchain model with a distributed learning technique, enables researchers to create a secure and immutable storage of computation history. The advantage of using blockchain together with distributed learning is that the master-server approach of conventional distributed learning is replaced by a decentralized architecture. Such an architecture defines the relationship between the partners in the network, without requiring that one trusted server mediates the work.

The use of blockchain for distributed learning has been proposed in recent works [19,20], however these studies only provide a proof-of-concept, without a fully decentralized solution supported by a blockchain platform. Furthermore, the scalability and privacy of this approach have yet to be evaluated [18].

In this work, we address these concerns with a novel blockchain-based approach to trace data provenance and safeguard the distributed learning process. Using the NSCLC-Radiomics dataset first introduced by Aerts et al. [21] we aim to confirm our hypothesis that, not only, our new decentralized model performs equivalent to a centralized model, but also provides the additional guarantee of traceability for the actions performed by all centers. Additionally, we validate our solution to demonstrate the ability of the blockchain distributed learning to leverage modern machine and deep learning techniques (e.g., convolutional neural networks).

Our objective, using the NSCLC-Radiomics dataset, is not to improve the signature developed by Aerts et al. [21] but rather to prove the feasibility of a blockchain based distributed learning approach and to illustrate that the distribution of data over multiple data centers provides similar results to the standard centralized approach.

This article makes the following contributions: 1) defines a blockchain-based protocol for training AI models using a distributed architecture; 2) shows how to construct classification models with sequential training on local datasets; and 3) demonstrates that the resulting blockchain model performs with comparable performance to that of a model where the training is conducted in centralized settings.

Background and significance

Learning from medical imaging
The process of extracting meaningful insights from medical images can be performed by applying Artificial Intelligence (AI) algorithms (i.e., machine learning or deep learning) [22]. Deep learning is a set of data decomposition and correlation algorithms inspired by similar processes within the human brain. These algorithms have been applied in

multiple fields including healthcare and medicine. convolutional neural networks (CNN), a class of deep learning, are commonly used to classify data from various data sources and medical images are no exception. AI algorithms are capable of extracting important information from medical images, which in turn, can be used in decision support systems to improve diagnostic, prognostic, or predictive accuracy [12,23].

Distributed learning

Distributed learning is a technique that supports multi-center machine learning, pioneered in 2013 [12]. These algorithms are designed to perform training while data remains in the local databases of each center [11,14,15,24]. The collaborators of a distributed learning process are connected to a master server that initializes and updates the learning. After initialization, each collaboration center trains a portion of the model on local data then provides the model weights to the master server. The master server in turn aggregates the weights, updates the model, and shares the updated model weights with the collaborators within the network. Each collaborator then retrains the local models based on the updated weights and sends them back to the master server to close the loop, which operates until a convergence threshold is reached, as illustrated in Figure 4.1. This approach, in principle, enables large-scale data/learning access, which improve performance and increase accuracy. In addition, distributed learning resolves legal and ethical privacy concerns associated with medical data by ensuring that sensitive data never leaves the firewalls of the medical centers.
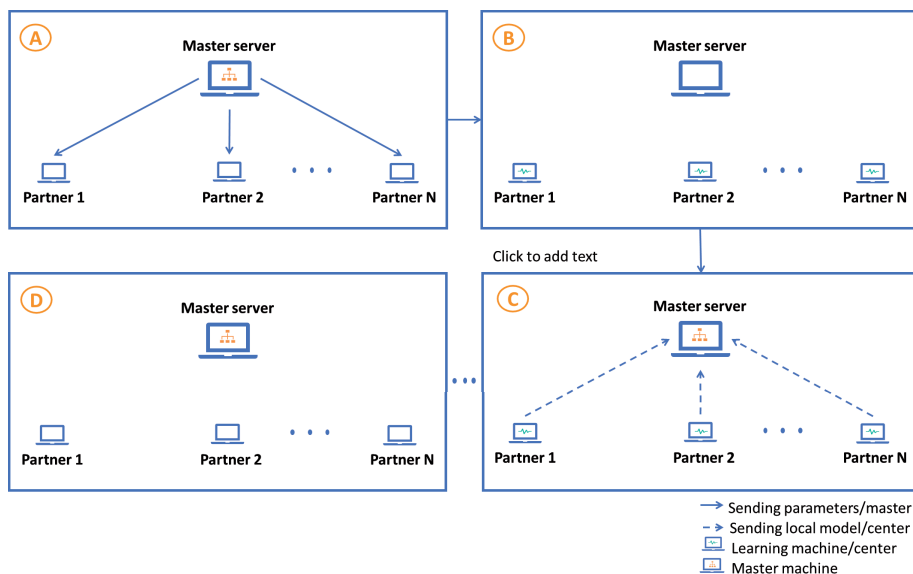

Figure 4.1: Conventional distributed learning (federated learning) process.
(A) Master server initialize the learning by sending initial models to the partners **(B)** partners train the received model with local data, **(C)** partners send updated models to the master se,rver, **(D)** master server aggregates the received models and verify convergence criteria.

Blockchain

Blockchain is a peer-to-peer (P2P) computational framework introduced in 2008 [25]. Transactions in a blockchain can be thought of as computational interactions between participants (such as the medical centers). Within a blockchain network, every participant can view and add interactions, but never modify the existing ones. This is due to the fact that interactions are stored in blocks, which are validated in the blockchain network. Each validated block contains a cryptographic hash of the previous block, thus making it impossible to forge interaction history in the system.

After its successful application within the cryptocurrency domain [25], blockchain technology subsequently received significant attention from the scientific community. This initial success instigated the use of blockchain in healthcare. Blockchains can now be used to ensure secure data sharing [26], compliance with license terms [27,28], drug counterfeiting prevention [29], amongst other applications in healthcare [30] and other domains [31].

Blockchain works via two regulating elements: a P2P network and a consensus protocol. The P2P network initiates and appends blocks representing the computations of the network. The consensus mechanism consists of a set of rules determining the contribution of each partner when validating the computations. A smart contract is a protocol that runs aside with the blockchain and enforces the rights and responsibilities of the network partners [32,33]. Once deployed, participants in the blockchain network can interact via smart contracts.

Materials and methods

Data

We used the open NSCLC-Radiomics dataset [34,35] to demonstrate this proof-of-concept study. The dataset consists of CT scans of 422 Non-Small Cell Lung Cancer (NSCLC) patients, paired with Gross Tumor Volume (GTV) segmentations (performed by an experienced radiologist), and the clinical outcome (survival). A summary of cohort and tumor specificities of the NSCLC-Radiomics dataset is presented in Table 4.1.

The generalizability of the proposed infrastructure was validated using the IRIS open dataset [36].

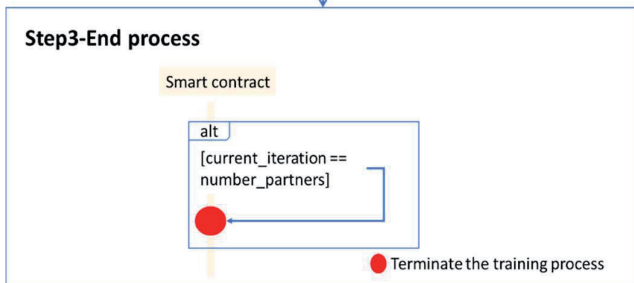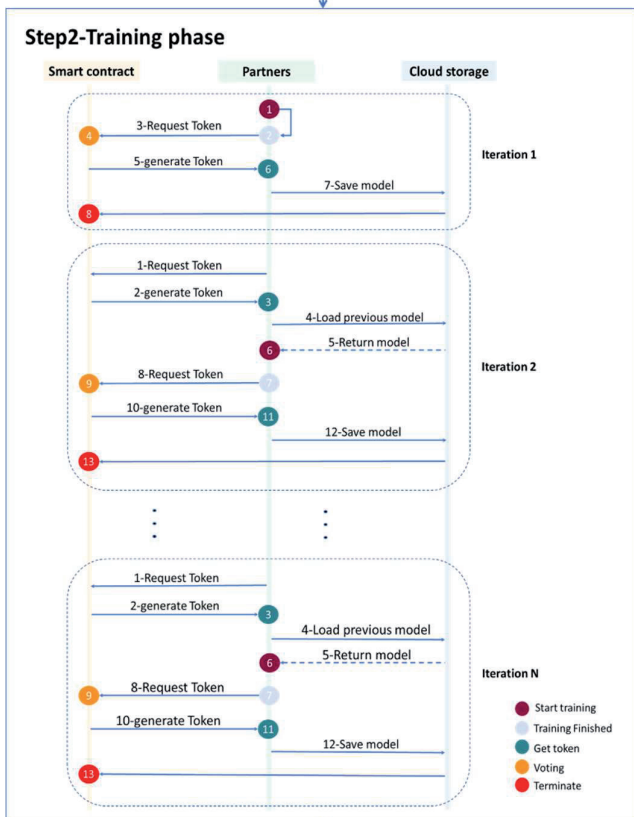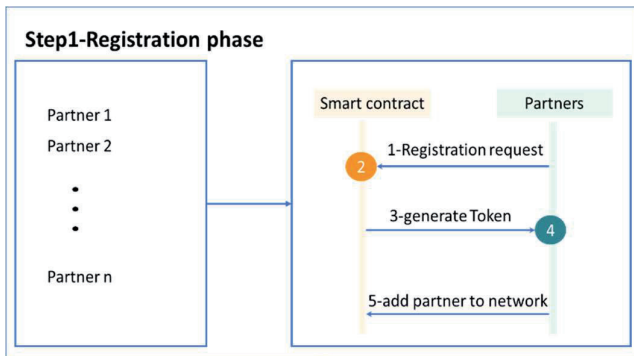**Table 4.1: Patient and tumor characteristics**

Variable	Frequency
Disease	
NSCLC	422
Gender	
Male	68.7 %
Female	32.3 %
TNM staging	
I	22.0 %
II	9.5 %
IIIa	26.5 %
IIIb	41.9 %
Treatment	
Radiotherapy	46.5 %
Radio-Chemotherapy	53.5%

Chained Distributed Machine learning (C-DistriM)

- **Architecture**

The objective of C-DistriM is to train distributed models with equivalent performance as centralized models, preserve data privacy, and increase trust amongst participating partners. C-DistriM leverages trust between the partners via the blockchain that stores unfalsifiable records of the training process. Figure 4.2 presents the overall architecture of C-DistriM. The smart contract of C-DistriM ensures:

- Creation of an organization structure representing the network of partners: the network of partners is stored within a smart contract. Each partner will take part in the learning process without moving the data to a trusted server.
- Confirmation of model deployment: saving each iteration of the model to cloud is considered as a new transaction in the blockchain. This requires that a consensus and an agreement on the current state of the blockchain to be reached by majority of the partners prior to appending the new transaction to the blockchain. Herein, every time a partner locally trains a model a majority must approve for the model to be saved in the cloud.
- Association of every partner with data quality and quantity: before confirming the model deployment, the transaction block with model accuracy statistics is established. This information determines if the contributions of the previous collaborator improved or negatively affected the model performance.
- Confirmation of model fetching: similarly, to "confirmation of model deployment", each time a partner requests a model from the cloud for subsequent training the majority of partners in the network must approve before the model to be downloaded.

Step1-Registration phase

Partner 1
Partner 2
•
•
•
Partner n

Smart contract   Partners

1-Registration request

3-generate Token

5-add partner to network

Step2-Training phase

Smart contract   Partners   Cloud storage

3-Request Token
5-generate Token
7-Save model
Iteration 1

1-Request Token
2-generate Token
4-Load previous model
5-Return model
8-Request Token
10-generate Token
12-Save model
Iteration 2

1-Request Token
2-generate Token
4-Load previous model
5-Return model
8-Request Token
10-generate Token
12-Save model
Iteration N

Start training
Training Finished
Get token
Voting
Terminate

Step3-End process

Smart contract

alt

[current_iteration ==
number_partners]

Terminate the training process

- Traceability of model leakage and data provenance. As all training records are saved to an append only chain in a timely manner, every model is linked to all partners that used it during the training process (load, upload, update). Similarly, every model can be linked to the data used to train/update it while maintaining data privacy concerns.

- **Implementation**

This work leveraged Ethereum blockchain [37], an open source smart contract platform, integrated with our distributed learning pipeline. We have implemented the smart contract using Solidity (compiler version 0.5.5), on RemiX IDE [38]. Solidity is an object-oriented programming language commonly used to implement smart contracts within the blockchain community.

Figure 4.3 illustrates the proposed distributed learning architecture. During each training iteration, a partner receives a token from the smart contract to start training the local model on local data. Once the model is trained the partner sends a request to archive the model to the cloud. Once approved by the majority, the smart contract returns a token allowing the partner to push the model to the cloud. Automated voting was performed based on the area under the receiver operating characteristic curve (AUC) of the model. If the model AUC deteriorates, a negative vote is cast, while improvements/no change in the AUC result in positive votes. In this prototype we used Google Cloud Storage (GCS) platform to store the shared models, where client-cloud communication was facilitated using the Python google-cloud-storage library (version 1.21.0). The models were encrypted and decrypted when being saved to or downloaded from the cloud. The encryption and decryption processes were performed using the Advanced Encryption Standard (AES-256) [37], as it is recommended for long term storage [39].

Ethereum is a public blockchain, implying that the C-DistriM computation history can be reviewed by the participating partners as well as the broader public. The prototype used the Ganache network [40], which allowed us to recreate the Ethereum blockchain

4

Figure 4.2: Overview of C-DistriM.
(Step 1) partners register to the network through the smart contract; (step 2) training starts by iterating through the partner list: (1) start training the first local model; (2) when training ends; (3) request a token to save the model to cloud; (4) vote to decide if model will be saved to cloud; (5) smart contract generates a token; (6) the partner gets the token; (7) and saves the model to cloud; (8) then the training is terminated for this partner. In the next iteration: (1) the partner will request a token to load the previous model from cloud; (2) smart contract generates a token; (3) the partner gets the token; (4) load previous model from cloud; (5) gets model; (6) start training; (7) when training ends; (8) request a token to save the model to cloud; (9) vote to decide if model will be saved to cloud; (10) smart contract generates a token; (11) the partner gets the token; (12) and saves the model to cloud; (13) then the training is terminated for this partner. The same process repeats for all partners. Step (3) if all partners finished training, then the training process is terminated.

platform for testing purposes. This means that our prototype is currently tested with a local test-net, however, the model is ready to be deployed in the public Ethereum.

While blockchain technology provides an auditable, traceable, and unfalsifiable structure to record distributed learning flow, it does not secure the learning process. To prevent any intentional or unintentional misuse of the downloaded model weights, by any of the partners, the model weight vectors were locked using the python portalocker library (version 1.5.2). During training, the training was initialized in the first iteration. The output model of the first iteration was used as a starting point for the next iteration, so is the new model. This process was repeated until all partners sequentially finish training. The last model in the queue was designated as the final output of the distributed learning process.
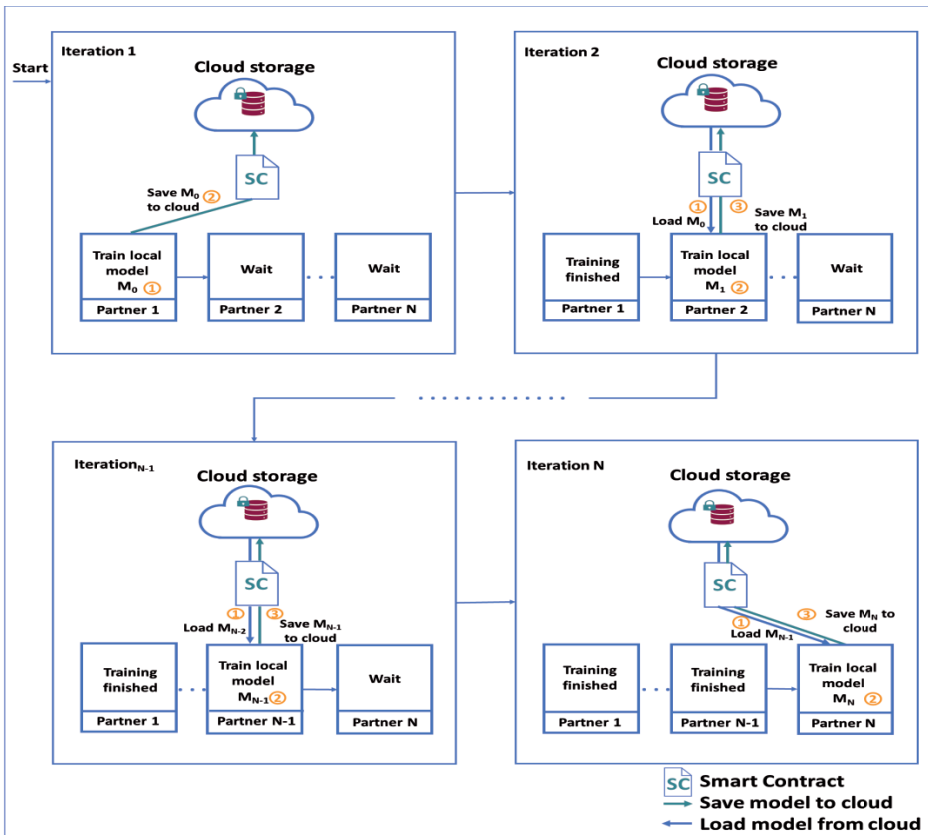
Figure 4.3: Distributed learning flow diagram, each iteration corresponds to one partner update of the model with local data.

Training

- **Data preparation**

Data augmentation was performed to balance the two classes (survive and not survive at a threshold of 2 years after start of treatment). The augmentation was performed in a different manner for each class: (1) the minority class in the training dataset was balanced by supplementing with zoom scaled variants of the images. After augmentation, the number of cases increased from 422 cases to 704 cases, (2) the images corresponding to the class that is represented high, non- survived (labeled 0) in the case of NSCLC-Radiomics dataset, was randomly augmented during the run-time (i.e., during training).

The data (n=704) was randomly split into training (n=563) and testing (n=141) sets (80% training and 20% testing) to train and evaluate the centralized training. Six testing scenarios were devised to validate the distributed infrastructure:

- "Scenario 1": simulation of a network of two partners by splitting the training data (same training data used to train the centralized model) into two subsets (n=281, and n=282 respectively).
- "Scenario 2": simulation of a network of three partners by splitting the training data into three subsets (n=188, n=189, and n=186 respectively).
- "Scenario 3": simulation of a network of four partners by splitting the training data into four subsets (n=141, n=140, n=141, and n=141 respectively).
- "Scenario 4": simulation of a coalition of two partners by splitting the training data into two non-equally distributed subsets (n=112, and n=451 respectively).
- "Scenario 5": simulation of a coalition of three partners by splitting the training data into three non-equally distributed subsets (n=113, n=67, and n=383 respectively).
- "Scenario 6": simulation of a coalition of four partners by splitting the training data into four non-equally distributed subsets (n=57, n=355, n=113, and n=38 respectively).

In all scenarios the models were evaluated using the same test data (n=141).

Data allocations were performed using scikit-learn library (version 0.22), therefore each partner in the training cycle held a balanced dataset. Once the data is prepared, they are split between the centers and run locally with the overall distributed learning process mediated by the C-DistriM blockchain model.

Data splits performed for the IRIS dataset are detailed in Supplementary materials (Section A).

4

- **Centralized model**

A previously validated 3D CNN binary classifier for two-year survival classification was implemented [41,42]. The CNN model is based on ResNet-18 [41]. The model consists of an input layer of shape (120, 160, 16), followed by 3x3 convolutional layers (while each convolutional layer is followed by a ReLu activation and batch normalization) with residual connections, the total number of convolutional layers is 18, in addition to an output layer entailing a sigmoid activation function. GTV segmentations were used to determine axial slices containing the tumor and crop them for training. As every GTV is of a different size, all cropped volumes were resized to (120 x 160 x 16) pixels for model training and validation process.

- **Decentralized models**

For each C-DistriM scenario, the batch size and validation-steps were adapted according to the number of data points in every center. The performance of both distributed and centralized models was quantified as the AUC of the receiver operating characteristic curve (ROC) and calibration curves. AUC values ranged from zero to one and the closest to one the AUC is, the better the model is. A calibration curve (or reliability curve) was defined as a plot of the relative frequency of empirical probability versus the predicted probability frequency. Calibration curves of ideal/optimized classifiers should fall close to the diagonal, as the estimated probabilities and empirical probabilities reach convergence.

Results

The model trained in a centralized approach, where all the data are contained in a single database and the training was performed without blockchain integration was used as the reference standard. We assessed the two-year prediction performance from the distributed and centralized survival CNN models, respectively. Table 4.2 summarizes the comparative performance in the test set of each approach (in 95% confidence interval).

A DeLong test [43] was used to compare the ROC curves and calculate the p-values to determine the differentiation significance between two independent means. The tests yielded a p-value of 0.102, 0.907, 0.984, 0.962, 0.747, and 0.779 when comparing the centralized model versus scenarios 1-6 respectively. The comparison of the ROC curves indicated that there is no statistically significant difference (all p-values > 0.05) between the performance of the distributed models and the centralized model in terms of discrimination, as shown in Figure 4.4. These results indicate that the distributed models can learn appropriate features in a comparable way to the centralized model learning; and that integration of distributed learning and blockchain is feasible.

Table 4.2: Discrimination performance (AUC) obtained by training centralized and distributed CNNs predicting 2-year NSCLC survival

Training typE	AUC (95% CI)
Centralized	0.76 (0.68-0.84)
Distributed (Scenario 1)	0.76 (0.73-0.88)
Distributed (Scenario 2)	0.76 (0.68-0.84)
Distributed (Scenario 3)	0.76 (0.68-0.84)
Distributed (Scenario 4)	0.76 (0.68-0.84)
Distributed (Scenario 5)	0.76 (0.68-0.83)
Distributed (Scenario 6)	0.77 (0.69-0.85)
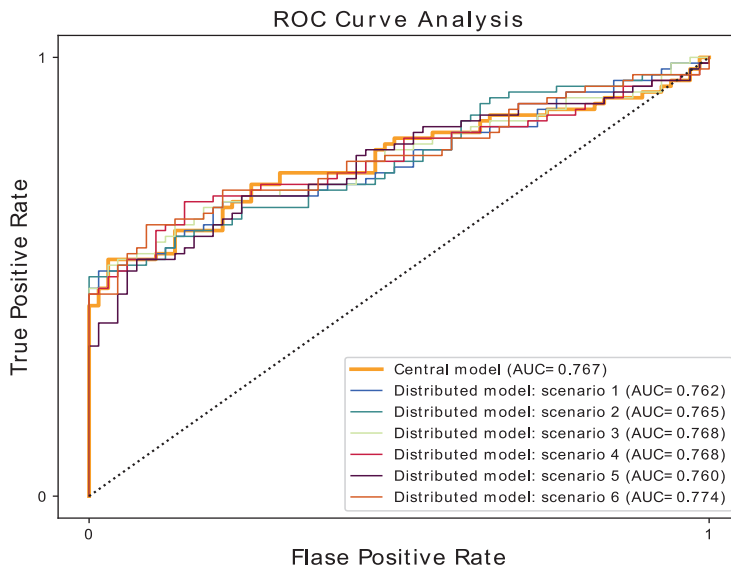
Figure 4.4: Receiver operator characteristic curves for two-year survival model trained using centralized learning and distributed learning.

From the calibration plots presented in Figure 4.5, we can observe a variation in the calibration of the models.

The IRIS conclusions were the same as the NSCLC-Radiomics dataset use-case. Detailed results are presented in Supplementary materials (Section A).

Supplementary materials (Section B) represents the ROC curves of each iteration of the scenario 3. The curves demonstrate how the learning improves when centers with more data are included in the training process.
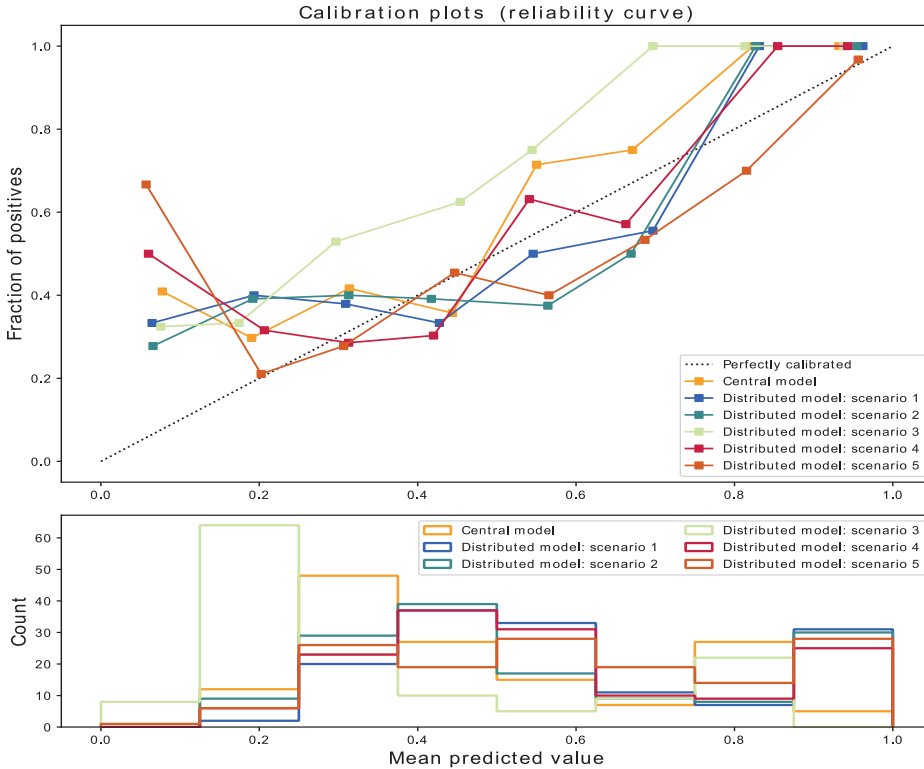
Figure 4.5: Calibration curves for centralized and distributed learning models.

Discussion and future work

Since its conception in 2013, Distributed learning has shown significant efficacy when leveraging big data to drive clinical insights [12]. This was recently demonstrated by Deist et al. who leveraged over 23,000 datapoints to train and validate a distributed logistic regression model, predicting post-treatment two-year survival [24]. In parallel researchers have developed methods to improve model generalizability [44], and promote training transparency via blockchain technologies [19,45]. Chen et al. proposed a fully blockchain-based privacy preserving distributed deep learning pipeline [19] where local model weights, from partners over the distributed network, are archived into the blockchain ledger as a transaction before being updated by the next collaborator's local data iteratively. Similar works have been demonstrated by Kuo et al. [45,46] leveraging blockchain using Logistic Regression machine learning models. While these pipelines [13,32,33] permit to secure local training and guarantee full traceability of the shared model weights, these methods are susceptible to drawbacks. These methods primarily employ fully visible model weights, which facilitate opportunities for misuse. Moreover, archiving the local weights to blockchain blocks along each iteration/partner

is costly (Ethers) and computationally expensive, and not recommended for a highly scalable system with a focus on throughput and efficiency.

Weng et al. proposed DeepChain, an optimized blockchain for secure distributed deep learning training [20], however the weights are saved directly to the blockchain and are accessible by all the partners within the distributed network. To overcome the risk of exposing the model weights, Lugan et al. [47] proposed to train distributed learning models on encrypted data, preventing any exposure of local weights. Nevertheless, when implementing deep learning and encrypting model weights, model design requires careful consideration as aspects such as the CNN activation functions must be adapted [48]. Model design challenges are exacerbated with the need for extensive computation power associated with complex encryption computations. Other studies proposed adding noise to the shared model weights [49,50], as an attempt to prevent the extraction of sensitive information. However, this approach can result in degradation of model performance. In this work, we proposed a solution to address the current challenges of distributed learning by means of blockchain and architectural modifications to the conventional distributed learning scheme. This work builds on previous applications by incorporating Ethereum, a validated, commercially used blockchain technology as opposed to ad hoc blockchain infrastructures[4]. The proposed approach, C-DistriM, secures the shared models within the distributed network by locking them when temporarily downloaded to local machines for process – alleviating concern of unauthorized use of models (i.e., edit, retrain, load model weights, or perform predictions). Post-training, C-DistriM: (1) encrypts the locally trained model, (2) uploads the encrypted model to the cloud and (3) removes all local copies preventing unauthorized exposure of the model. In contrary to other solutions, such as training on encrypted data, C-Distrim: maintains the native implementations of machine/deep learning algorithms that may be used for training. We observed that the AUC for distributed learning models generated by C-DistriM do not differ with statistical significance from models trained in classical centralized configuration. Calibration plots between models indicated a slight variation between the predicted scores. As CNN models have hundreds of millions of parameters that may influence stable performance dependent on (1) the size and type of training data, and (2) optimal batch size. Additionally, the last layer of a CNN is not in the proper scale to evaluate the reliability of the model [51]. To obtain appropriate probabilities, one may consider rescaling the predictions by applying Platt Scaling [52] or Temperature Scaling [53], however this was out of the scope of this work.

While blockchain infrastructure does permit archiving model iterations within the blockchain ledger, blockchains are not suitable for large data storage [37]. To mitigate this

4

---

4   Ad hoc blockchains refer to use implementations that are designed to replicate blockchains for test purposes but are not suitable for deployment.

concern, C-Distrim archives model iterations over the cloud, while the blockchain is used exclusively to store model metadata (i.e., partner name, and model name – composed of the partner name and iteration number) and monitor the training performance. Based on the performance of the model in a particular iteration, the improvement or deterioration of the model can be traced back to a particular dataset/partner. Blockchain tokens are used to generate access permissions to the model in the cloud. C-Distrim facilitates the ability to "restore" a prior model state and retrain an updated model by skipping the training step for a particular partner in the case of model performance degradation. This functionality can also be used as an internal quality control metric to flag the incorporation of poor data into the training cycle. Table 4.3 illustrates the key differences between the listed blockchain infrastructures and the proposed C-DistriM.

Table 4.3: Summary of different blockchain infrastructures used for privacy preserving distributed learning

Ref	Blockchain	Shared information	Smart contract
[19]	Public	• Model gradient	• Authentication
[20]	Permissioned	• Model gradient • Data size • Data format • Data topic • Data description	• Coordinates computation
[46]	Permissioned	• Model mean • Model covariance • Model type • Accuracy • Iteration number	Not applicable
[47]	Experimented with both public & private	• Hash of the model's parameters • Participant ID • Participant strength	Not applicable
C-DistriM	Public	• Token of model's cloud location • Participant ID • AUC • Iteration number • Approval of trained models • The last updated model (final model)	• Coordinates access to consortium • Coordinates model training • Coordinates the partners and access to the trained models in the cloud

One of the key features of the C-DistriM infrastructure is its traceability. Traceability of the data and lineage of the AI algorithms are key components of trustworthy AI. As the transaction records are immutable on blockchain ledgers, we can trace back any action performed by any of the participating partners at any time. Furthermore, due

to the inherent traceability of our infrastructure, it is expected that all the participating partners will have accentuated trust in using the process. The blockchain ledger can also foster commercial discussions such as royalties for the new AI algorithms proportional to the number of patients provided by each partner in the distributed network.

Deploying a smart contract to public Ethereum is payable. Thereby, it is important to note that this work was developed using the development and testing environments provided by Ethereum.

In future works, we will extend the C-DistriM pipeline to monitor the applications of the final models and integrate a web portal accessible by all the participating partners to visualize the transaction history. We also intend to extend our development cycle using the Ethereum test networks to simulate a real-word distributed learning network and measure its performance in terms of scalability and costs. Finally, we wish to investigate how C-DistriM performs when malicious partners are intentionally added to the network.

Conclusion

In this work, we validated our hypothesis which is Chained Distributed Machine learning combined with a blockchain-based platform (C-DistriM), is feasible and gives a similar result to the traditional centralized approach. Furthermore, the blockchain architecture was beneficial to trace data origin and monitor the training process against model degradation and dishonest behaviors. We believe this approach will increase trust between parties therefore stimulate collaboration globally between parties when delivering robust AI informed by big data.

References

[1]   S. Kulkarni, N. Seneviratne, M.S. Baig, A.H.A. Khan, Artificial Intelligence in Medicine: Where Are We Now?, Academic Radiology. 27 (2020) 62–70. https://doi.org/10.1016/j.acra.2019.10.001.

[2]   Y. Zhou, J. Xu, Q. Liu, C. Li, Z. Liu, M. Wang, H. Zheng, S. Wang, A Radiomics Approach With CNN for Shear-Wave Elastography Breast Tumor Classification, IEEE Trans. Biomed. Eng. 65 (2018) 1935–1942. https://doi.org/10.1109/TBME.2018.2844188.

[3]   X. Jiang, J. Li, Y. Kan, T. Yu, S. Chang, X. Sha, H. Zheng, Y. Luo, S. Wang, MRI Based Radiomics Approach with Deep Learning for Prediction of Vessel Invasion in Early-Stage Cervical Cancer, IEEE/ACM Trans. Comput. Biol. and Bioinf. (2020) 1–1. https://doi.org/10.1109/TCBB.2019.2963867.

[4]   P. Lambin, R.T.H. Leijenaar, T.M. Deist, J. Peerlings, E.E.C. de Jong, J. van Timmeren, S. Sanduleanu, R.T.H.M. Larue, A.J.G. Even, A. Jochems, Y. van Wijk, H. Woodruff, J. van Soest, T. Lustberg, E. Roelofs, W. van Elmpt, A. Dekker, F.M. Mottaghy, J.E. Wildberger, S. Walsh, Radiomics: the bridge between medical imaging and personalized medicine, Nat Rev Clin Oncol. 14 (2017) 749–762. https://doi.org/10.1038/nrclinonc.2017.141.

[5]   M. Owais, M. Arsalan, J. Choi, K.R. Park, Effective Diagnosis and Treatment through Content-Based Medical Image Retrieval (CBMIR) by Using Artificial Intelligence, JCM. 8 (2019) 462. https://doi.org/10.3390/jcm8040462.

[6]   S. Guo, Z. Yang, Multi-Channel-ResNet: An integration framework towards skin lesion analysis, Informatics in Medicine Unlocked. 12 (2018) 67–74. https://doi.org/10.1016/j.imu.2018.06.006.

[7]   W. Hua, T. Xiao, X. Jiang, Z. Liu, M. Wang, H. Zheng, S. Wang, Lymph-vascular space invasion prediction in cervical cancer: Exploring radiomics and deep learning multilevel features of tumor and peritumor tissue on multiparametric MRI, Biomedical Signal Processing and Control. 58 (2020) 101869. https://doi.org/10.1016/j.bspc.2020.101869.

[8]   S. Walsh, E.E.C. de Jong, J.E. van Timmeren, A. Ibrahim, I. Compter, J. Peerlings, S. Sanduleanu, T. Refaee, S. Keek, R.T.H.M. Larue, Y. van Wijk, A.J.G. Even, A. Jochems, M.S. Barakat, R.T.H. Leijenaar, P. Lambin, Decision Support Systems in Oncology, JCO Clinical Cancer Informatics. (2019) 1–9. https://doi.org/10.1200/CCI.18.00001.

[9]   C. Garling, Andrew Ng: Why 'Deep Learning' Is a Mandate for Humans, Not Just Machines, Wired. (2015). https://www.wired.com/brandlab/2015/05/andrew-ng-deep-learning-mandate-humans-not-just-machines/ (accessed June 11, 2019).

[10]   R. Sullivan, Delivering affordable cancer care in high-income countries, The Lancet Oncology. 12 (2011) 48. https://doi.org/10.1016/S1470-2045(11)70141-3.

[11]   A. Jochems, T.M. Deist, J. van Soest, M. Eble, P. Bulens, P. Coucke, W. Dries, P. Lambin, A. Dekker, Distributed learning: Developing a predictive model based on data from multiple hospitals without data leaving the hospital – A real life proof of concept, Radiotherapy and Oncology. 121 (2016) 459–467. https://doi.org/10.1016/j.radonc.2016.10.002.

[12]   P. Lambin, E. Roelofs, B. Reymen, E.R. Velazquez, J. Buijsen, C.M.L. Zegers, S. Carvalho, R.T.H. Leijenaar, G. Nalbantov, C. Oberije, M. Scott Marshall, F. Hoebers, E.G.C. Troost, R.G.P.M. van Stiphout, W. van Elmpt, T. van der Weijden, L. Boersma, V. Valentini, A. Dekker, 'Rapid Learning health care in oncology' – An approach towards decision support systems enabling customised radiotherapy', Radiotherapy and Oncology. 109 (2013) 159–164. https://doi.org/10.1016/j.radonc.2013.07.007.

[13]   F. Zerka, S. Barakat, S. Walsh, M. Bogowicz, R.T.H. Leijenaar, A. Jochems, B. Miraglio, D. Townend, P. Lambin, Systematic Review of Privacy-Preserving Distributed Machine Learning From Federated Databases in Health Care, JCO Clinical Cancer Informatics. (2020) 184–200. https://doi.org/10.1200/CCI.19.00047.

[14]   T.M. Deist, A. Jochems, J. van Soest, G. Nalbantov, C. Oberije, S. Walsh, M. Eble, P. Bulens, P. Coucke, W. Dries, A. Dekker, P. Lambin, Infrastructure and distributed learning methodology for privacy-preserving multi-centric rapid learning health care: euroCAT, Clinical and Translational Radiation Oncology. 4 (2017) 24–31. https://doi.org/10.1016/j.ctro.2016.12.004.

[15]   T.M. Deist, F.J.W.M. Dankers, G. Valdes, R. Wijsman, I.-C. Hsu, C. Oberije, T. Lustberg, J. van Soest, F. Hoebers, A. Jochems, I. El Naqa, L. Wee, O. Morin, D.R. Raleigh, W. Bots, J.H. Kaanders, J. Belderbos, M. Kwint, T. Solberg, R. Monshouwer, J. Bussink, A. Dekker, P. Lambin, Machine learning algorithms for outcome prediction in (chemo)radiotherapy: An empirical comparison of classifiers, Medical Physics. 45 (2018) 3449–3459. https://doi.org/10.1002/mp.12967.

[16]   Z. Shi, I. Zhovannik, A. Traverso, F.J.W.M. Dankers, T.M. Deist, P. Kalendralis, R. Monshouwer, J. Bussink, R. Fijten, H.J.W.L. Aerts, A. Dekker, L. Wee, Distributed radiomics as a signature validation study using the Personal Health Train infrastructure, Sci Data. 6 (2019) 218. https://doi.org/10.1038/s41597-019-0241-0.

[17]   M.J. Sheller, G.A. Reina, B. Edwards, J. Martin, S. Bakas, Multi-institutional Deep Learning Modeling Without Sharing Patient Data: A Feasibility Study on Brain Tumor Segmentation, in: A. Crimi, S. Bakas, H. Kuijf, F. Keyvan, M. Reyes, T. van Walsum (Eds.), Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries, Springer International Publishing, Cham, 2019: pp. 92–104. https://doi.org/10.1007/978-3-030-11723-8_9.

[18]   C. Song, T. Ristenpart, V. Shmatikov, Machine Learning Models that Remember Too Much, in: Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security, ACM, Dallas Texas USA, 2017: pp. 587–601. https://doi.org/10.1145/3133956.3134077.

[19]   X. Chen, J. Ji, C. Luo, W. Liao, P. Li, When Machine Learning Meets Blockchain: A Decentralized, Privacy-preserving and Secure Design, in: 2018 IEEE International Conference on Big Data (Big Data), IEEE, Seattle, WA, USA, 2018: pp. 1178–1187. https://doi.org/10.1109/BigData.2018.8622598.

[20]   J. Weng, J. Weng, J. Zhang, M. Li, Y. Zhang, W. Luo, DeepChain: Auditable and Privacy-Preserving Deep Learning with Blockchain-based Incentive, IEEE Trans. Dependable and Secure Comput. (2019) 1–1. https://doi.org/10.1109/TDSC.2019.2952332.

[21]   H.J.W.L. Aerts, E.R. Velazquez, R.T.H. Leijenaar, C. Parmar, P. Grossmann, S. Carvalho, J. Bussink, R. Monshouwer, B. Haibe-Kains, D. Rietveld, F. Hoebers, M.M. Rietbergen, C.R. Leemans, A. Dekker, J. Quackenbush, R.J. Gillies, P. Lambin, Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach, Nat Commun. 5 (2014) 4006. https://doi.org/10.1038/ncomms5006.

[22]   T.M. Mitchell, Machine learning, International ed., [Reprint.], McGraw-Hill, New York, NY, 1997.

[23]   K. Kourou, T.P. Exarchos, K.P. Exarchos, M.V. Karamouzis, D.I. Fotiadis, Machine learning applications in cancer prognosis and prediction, Computational and Structural Biotechnology Journal. 13 (2015) 8–17. https://doi.org/10.1016/j.csbj.2014.11.005.

[24]   T.M. Deist, F.J.W.M. Dankers, P. Ojha, M. Scott Marshall, T. Janssen, C. Faivre-Finn, C. Masciocchi, V. Valentini, J. Wang, J. Chen, Z. Zhang, E. Spezi, M. Button, J. Jan Nuyttens, R. Vernhout, J. van Soest, A. Jochems, R. Monshouwer, J. Bussink, G. Price, P. Lambin, A. Dekker, Distributed learning on 20 000+ lung cancer patients – The Personal Health Train, Radiotherapy and Oncology. 144 (2020) 189–200. https://doi.org/10.1016/j.radonc.2019.11.019.

[25]   S. Nakamoto, Bitcoin: A Peer-to-Peer Electronic Cash System, (n.d.) 9.

[26]   P. Zhang, J. White, D.C. Schmidt, G. Lenz, S.T. Rosenbloom, FHIRChain: Applying Blockchain to Securely and Scalably Share Clinical Data, Computational and Structural Biotechnology Journal. 16 (2018) 267–278. https://doi.org/10.1016/j.csbj.2018.07.004.

[27]   N. Havelange, M. Dumontier, B. Wouters, J. Linde, D. Townend, A. Riedl, V. Urovi, LUCE: A Blockchain Solution for monitoring data License accoUntability and CompliancE, ArXiv:1908.02287 [Cs]. (2019). http://arxiv.org/abs/1908.02287 (accessed December 27, 2019).

[28]   V. Jaiman, V. Urovi, A Consent Model for Blockchain-based Health Data Sharing Platforms, IEEE Access. (2020). https://doi.org/10.1109/ACCESS.2020.3014565.

[29]   S. Vruddhula, Application of on-dose identification and blockchain to prevent drug counterfeiting, Pathogens and Global Health. 112 (2018) 161–161. https://doi.org/10.1080/20477724.2018.1503268.

[30]   T.-T. Kuo, H.-E. Kim, L. Ohno-Machado, Blockchain distributed ledger technologies for biomedical and health care applications, Journal of the American Medical Informatics Association. 24 (2017) 1211–1220. https://doi.org/10.1093/jamia/ocx068.

[31]   H.-N. Dai, Z. Zheng, Y. Zhang, Blockchain for Internet of Things: A Survey, IEEE Internet Things J. 6 (2019) 8076–8094. https://doi.org/10.1109/JIOT.2019.2920987.

[32]   C.D. Clack, V.A. Bakshi, L. Braine, Smart Contract Templates: foundations, design landscape and research directions, ArXiv:1608.00771 [Cs]. (2017). http://arxiv.org/abs/1608.00771 (accessed January 27, 2020).

[33]   Z. Zheng, S. Xie, H.-N. Dai, W. Chen, X. Chen, J. Weng, M. Imran, An overview on smart contracts: Challenges, advances and platforms, Future Generation Computer Systems. 105 (2020) 475–491. https://doi.org/10.1016/j.future.2019.12.019.

4

[34] H.J.W.L. Aerts, E. Rios Velazquez, R.T.H. Leijenaar, P. Chintan, P. Grossmann, S. Carvalho, J. Bussink, R. Monshouwer, B. Haibe-Kains, D. Rietveld, F. Hoebers, M. M. Rietbergen, C.R. Leemans, A. Dekker, J. Quackenbush, R. Barkovich, P. Lambin, Data From NSCLC-Radiomics [Data set], (2019). https://doi.org/10.7937/K9/TCIA.2015.PF0M9REI.

[35] K. Clark, B. Vendt, K. Smith, J. Freymann, J. Kirby, P. Koppel, S. Moore, S. Phillips, D. Maffitt, M. Pringle, L. Tarbox, F. Prior, The Cancer Imaging Archive (TCIA): Maintaining and Operating a Public Information Repository, J Digit Imaging. 26 (2013) 1045–1057. https://doi.org/10.1007/s10278-013-9622-7.

[36] R.A. Fisher, The use of multiple measurements in taxonomic problems, Annals of Eugenics. 7 (1936) 179–188. https://doi.org/10.1111/j.1469-1809.1936.tb02137.x.

[37] D.G. Wood, ETHEREUM: a secure decentralised generalised transaction ledger, (n.d.) 32.

[38] Remix, https://remix.ethereum.org/, Ethereum, n.d. https://remix.ethereum.org/ (accessed August 14, 2020).

[39] E. Barker, A. Roginsky, Transitioning the use of cryptographic algorithms and key lengths, National Institute of Standards and Technology, Gaithersburg, MD, 2019. https://doi.org/10.6028/NIST.SP.800-131Ar2.

[40] ganache, https://www.trufflesuite.com/ganache, n.d. https://www.trufflesuite.com/ganache (accessed August 14, 2020).

[41] K. He, X. Zhang, S. Ren, J. Sun, Deep Residual Learning for Image Recognition, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, Las Vegas, NV, USA, 2016: pp. 770–778. https://doi.org/10.1109/CVPR.2016.90.

[42] C. Li, J. Xu, Q. Liu, Y. Zhou, L. Mou, Z. Pu, Y. Xia, H. Zheng, S. Wang, Multi-View Mammographic Density Classification by Dilated and Attention-Guided Residual Learning, IEEE/ACM Trans. Comput. Biol. and Bioinf. (2020) 1–1. https://doi.org/10.1109/TCBB.2020.2970713.

[43] E.R. DeLong, D.M. DeLong, D.L. Clarke-Pearson, Comparing the Areas under Two or More Correlated Receiver Operating Characteristic Curves: A Nonparametric Approach, Biometrics. 44 (1988) 837. https://doi.org/10.2307/2531595.

[44] X. Wu, J. Zhang, F.-Y. Wang, Stability-Based Generalization Analysis of Distributed Learning Algorithmsfor Big Data, IEEE Trans. Neural Netw. Learning Syst. (2019) 1–12. https://doi.org/10.1109/TNNLS.2019.2910188.

[45] T.-T. Kuo, R.A. Gabriel, L. Ohno-Machado, Fair compute loads enabled by blockchain: sharing models by alternating client and server roles, Journal of the American Medical Informatics Association. 26 (2019) 392–403. https://doi.org/10.1093/jamia/ocy180.

[46] T.-T. Kuo, J. Kim, R.A. Gabriel, Privacy-preserving model learning on a blockchain network-of-networks, Journal of the American Medical Informatics Association. 27 (2020) 343–354. https://doi.org/10.1093/jamia/ocz214.

[47] S. Lugan, P. Desbordes, E. Brion, L.X. Ramos Tormo, A. Legay, B. Macq, Secure Architectures Implementing Trusted Coalitions for Blockchained Distributed Learning (TCLearn), IEEE Access. 7 (2019) 181789–181799. https://doi.org/10.1109/ACCESS.2019.2959220.

[48] Q. Zhang, L.T. Yang, Z. Chen, Privacy Preserving Deep Computation Model on Cloud for Big Data Feature Learning, IEEE Trans. Comput. 65 (2016) 1351–1362. https://doi.org/10.1109/TC.2015.2470255.

[49] R. Shokri, V. Shmatikov, Privacy-Preserving Deep Learning, in: Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security - CCS '15, ACM Press, Denver, Colorado, USA, 2015: pp. 1310–1321. https://doi.org/10.1145/2810103.2813687.

[50] Y. Lu, X. Huang, Y. Dai, S. Maharjan, Y. Zhang, Blockchain and Federated Learning for Privacy-Preserved Data Sharing in Industrial IoT, IEEE Trans. Ind. Inf. 16 (2020) 4177–4186. https://doi.org/10.1109/TII.2019.2942190.

[51] Y. Li, C. Huang, L. Ding, Z. Li, Y. Pan, X. Gao, Deep learning in bioinformatics: Introduction, application, and perspective in the big data era, Methods. 166 (2019) 4–21. https://doi.org/10.1016/j.ymeth.2019.04.008.

[52] J. Platt, Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods, (1999) 61–74. https://doi.org/10.1.1.41.1639.

[53] C. Guo, G. Pleiss, Y. Sun, K.Q. Weinberger, On Calibration of Modern Neural Networks, (n.d.) 10.

Supplementary materials

Section A
Materials and methods

Data preparation

The IRIS dataset (n=150) contains three iris species. The dataset classes are balanced, fifty examples for each species, therefore we did not any preprocessing on the data.

We randomly split the data (n=150) into training (n=120) and testing (n=30) sets to train and evaluate the centralized training. Following scenarios were prepared and executed:

- "Scenario 1": a simulation of a network of two partners by splitting the training data (same training data used to train the centralized model) into two subsets (n=60, and n=60 respectively).
- "Scenario 2": a simulation of a network of three partners by splitting the training data into three subsets (n=40, n=40, and n=40 respectively).
- "Scenario 3": a simulation of a network of four partners by splitting the training data into four subsets (n=30, n=30, n=30, and n=30 respectively).
- "Scenario 4": we simulated a network of two partners by splitting the training data into two non-equally distributed subsets (n=80, and n=40 respectively).
- "Scenario 5": a simulation of a network of three partners by splitting the training data into three non-equally distributed subsets (n=24, n=57, and n=39 respectively).
- "Scenario 6": a simulation of a network of four partners by splitting the training data into four non-equally distributed subsets (n=24, n=39, n=18, and n=39 respectively).

In all scenarios the models were evaluated using the same test data (n=30).

All data splits were performed using scikit-learn library version 0.22.

Training

C-Distrim was prepared to train multi-class neural network. Once the data is prepared, they are split between the local centers and run locally using the C-DistriM infrastructure.

Results

The ROC curves for each class are illustrated in Figure S4.1, and the comparative performance of each approach (in 95% confidence interval) is illustrated in Table S4.1.
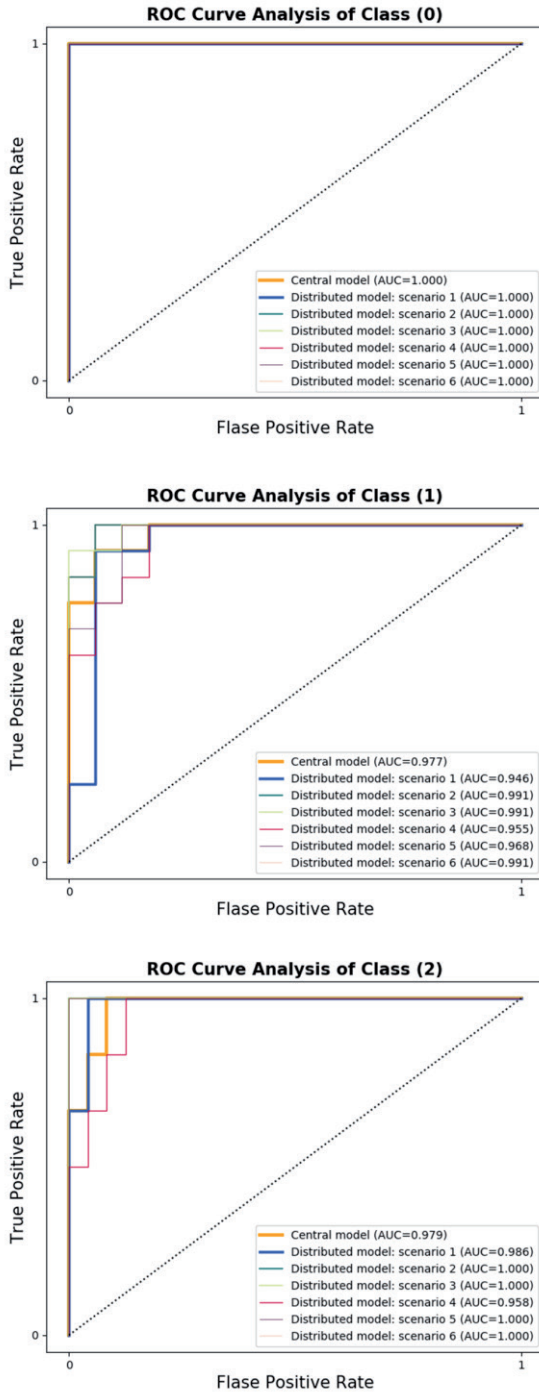
4

Figure S4.1: Receiver operator characteristic curves for IRIS species prediction model trained using centralized learning and distributed learning.

Table S4.1: Discrimination performance (AUC) obtained by training centralized and distributed multiclass neural network predicting iris species

Training type	AUC (95% CI)		
	Class 0	Class 1	Class 2
Centralized	1 (1-1)	0.97 (0.93-1)	0.97 (0.93-1)
Distributed (Scenario 1)	1 (1-1)	0.94 (0.85-1)	0.98 (0.95-1)
Distributed (Scenario 2)	1 (1-1)	0.99 (0.96-1)	1 (1-1)
Distributed (Scenario 3)	1 (1-1)	0.99 (0.96-1)	1 (1-1)
Distributed (Scenario 4)	1 (1-1)	0.95 (0.89-1)	0.95 (0.89-1)
Distributed (Scenario 5)	1 (1-1)	0.96 (0.91-1)	1 (1-1)
Distributed (Scenario 6)	1 (1-1)	0.99 (0.96-1)	1 (1-1)

The Delong tests yielded a *p*-value of 1 when comparing the centralized model versus scenarios 1-6 respectively for class 0. The tests yielded a *p*-value of 0.508, 0.544, 0.575, 0.310, 0.778, and 0.544 when comparing the centralized model versus scenarios 1-6 respectively for class 1. The tests yielded a *p*-value of 0.479, 0.318, 0.318, 0.309, 0.318, and 0.318 when comparing the centralized model versus scenarios 1-6 respectively for class 2.

Section B
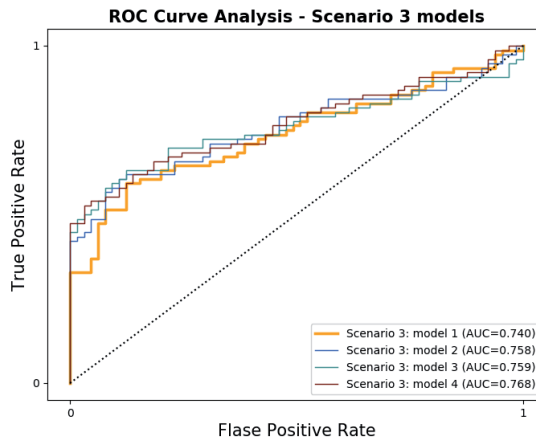
Figure S4.2: Receiver operator characteristic curves for two-year survival models trained using "Scenario 3" data distribution.
(A) model 1: data from center 1 only (n=141); **(B)** model 2: data from center 1 and center 2 (n=141 + n=140); **(C)** data from center 1, center 2, center 3 (n=141 + n=140 + n=141); **(D)** data from center 1, center 2, center 3, center 4 (n=141 + n=140 + n=141 + n=141).

Evaluation

5

Privacy preserving distributed learning classifiers – sequential learning with small sets of data

Fadila Zerka | Visara Urovi | Fabio Bottari | Ralph T.H. Leijenaar |
Sean Walsh | Hanif Gabrani-Juma | Martin Gueuning |
Akshayaa Vaidyanathan | Wim Vos | Mariaelena Occhipinti |
Henry C. Woodruff | Michel Dumontier | Philippe Lambin

Abstract

Artificial intelligence (AI) typically requires a significant amount of high-quality data to build reliable models, where gathering enough data within a single institution can be particularly challenging. In this study we investigated the impact of using sequential learning to exploit very small, siloed sets of clinical and imaging data to train AI models. Furthermore, we evaluated the capacity of such models to achieve equivalent performance when compared to models trained with the same data over a single centralized database.

We propose a privacy preserving distributed learning framework, learning sequentially from each dataset. The framework is applied to three machine learning algorithms: Logistic Regression, Support Vector Machines (SVM), and Perceptron. The models were evaluated using four open-source datasets (Breast cancer, Indian liver, NSCLC-Radiomics dataset, and Stage III NSCLC).

The proposed framework ensured a comparable predictive performance against a centralized learning approach. Pairwise DeLong tests showed no significant difference between the compared pairs for each dataset.

Distributed learning contributes to preserve medical data privacy. We foresee this technology will increase the number of collaborative opportunities to develop robust AI, becoming the default solution in scenarios where collecting enough data from a single reliable source is logistically impossible. Distributed sequential learning provides privacy persevering means for institutions with small but clinically valuable datasets to collaboratively train predictive AI while preserving the privacy of their patients. Such models perform similarly to models that are built on a larger central dataset.

Abstract

Introduction

The application of artificial intelligence (AI) (i.e., machine/ deep learning models) within the clinical decision making process, also referred to as precision medicine, has become a research topic of increasing interest [1,2]. The rising number of published AI models in the literature that support diagnosis/ prognosis is a testament to this.

The most common way to train AI models, often referred to as "centralized training", is when the data is sourced from a single centralized database and the training of the classification AI model is local to a single machine. This approach however is not ideal during collaborative efforts where data sharing and centralization is strictly regulated by legal and ethical considerations. For instance, the General Data Protection Regulation (GDPR) and Health Insurance Portability and Accountability Act (HIPAA) act as safeguards to protect the privacy of patient data. Distributed learning (i.e., federated learning, ensemble learning, or sequential learning) offers a promising solution to this centralization barrier, allowing development and validation of predictive models while preserving the privacy the patient data. Federated learning, the most conventional form of distributed learning, involves a master server that coordinates the initialization and aggregation of learning within a consortium of partners [3,4]. Ensemble learning consists of training independent models on local data, and each model's predictions on new data are grouped to a single global prediction [5]. Sequential distributed learning is an extension of distributed learning enabling the partners of a consortium to iteratively update a model with their respective local datasets. The last model in the queue is the final model [6,7]. These approaches are particularly appealing in the cases of small datasets (e.g., low clinical volume or rare diseases) in which the amount of data available to a single center is below the threshold to develop robust and generalizable AI. Since the performance and the robustness of an AI model is directly related to the number of samples on which it was trained and validated [2], the scarcity of data coupled with lengthy procedures required to centralize data can derail initiatives to develop clinical decision support tools.

While distributed learning has been well established with applications in multicentric studies [3,6,8–10], and previous work on ensemble distributed learning on small local datasets has indicated promising performance [11,12], the impact of the network data-scape (e.g., small batch sizes) has yet to be systematically investigated for sequential distributed learning. In this work, we investigate the performance of Stochastic Gradient Descent (SGD) based classifiers trained using a sequential distributed learning approach. We evaluate the influence on model performance when using micro batch sizes (as small as $n = 1$) to replicate cases where a participating institution (or partner) may only provide a single case record to the consortium to support training. To this extent we examine the influence of micro batch sizes on sequential learning model performance compared to the equivalent (i.e., same data) centralized model using a variety of Radiomics and clinical open-source datasets.

5

Background and significance

Model Optimizers - Stochastic gradient descent
This work also explores stochastic gradient descent (SGD) which is an iterative optimization method. It is a commonly used optimization technique applied to various machine and deep learning algorithms[13]. Upon each training iteration, the SGD optimizer fine-tunes the algorithm, minimizing the error of the model. As opposed to standard gradient descent optimizers, where the error is reduced over the entirety of the training dataset, SGD randomly selects small training batches and approximates the gradient for the random batch. The iterative process of batch selection is performed by randomly shuffling the dataset and minimizing over all batches, offering the advantage of avoiding local minima and reducing model optimization time.

Challenges in medical image analysis
Multicentric studies are needed to develop robust AI and to demonstrate the clinical relevance of imaging AI. This kind of studies face many challenges such as:

1. Data collection (described in section "Medical data sharing");

2. Data heterogeneity, caused by the difference in acquisition and reconstruction settings amongst the different medical centers[14]. To ensure better model building in a heterogeneous domain, the raw data and/or the features derived from it must be harmonized [15,16];

3. And Inter-reader variability, the automation of manual tasks, such as organ and lesion delineation, requires to learn from ground truth masks delineated manually by experienced radiologists [17]. The difference in experience and trainings of the clinicians leads to a variation on the ground truth delineations, which in turn represents a challenge in segmentation model training and validation [17].

Medical data sharing
Despite the efforts made to publicly share medical data in public repositories, including, the cancer imaging archive (TCIA; https://www.cancerimagingarchive.net/), and the NIH BioLINCC (https://biolincc.nhlbi.nih.gov/home/), among others [18], data sharing remains very difficult, especially in low prevalence rare diseases. Within the context of rare-diseases, data sharing limitations can hinder rare disease research and development, as well documented cases may be limited in number. This proves especially difficult in situations where a single institution may want to extract hidden insights using machine learning approaches, such as a diagnostic or prognostic biomarker. Initiatives, such as the European joint program on rare diseases (EJP RD; https://www.

ejprarediseases.org/index.php/about/), began to address this issue and has illustrated the potential of data in driving precision medicine and accelerating rare disease diagnosis/ prognosis.

The importance of datatype (e.g., genotype, phenotype and endotype among others) in modeling patients with rare diseases, is well demonstrated within the literature [19,20]. However, de-identification of patient data prior to sharing, does not necessarily guarantee preservation of privacy [21] as patient personal information can potentially be re-identified from the de-identified features (e.g., up to 99.98% of the American population in any dataset can be identified using only 15 demographic features) [22]. This risk increases as the dimensionality of data increases. In order to protect patient sensitive information, data acquisition and sharing is therefore tightly regulated by ethical and legal constraints [23]. In this context, distributed sequential learning is an important approach to facilitate data analysis across institutions while preserving data privacy.

Distributed learning
Distributed learning was first applied to clinical decision support systems in 2013 [2]. Distributed learning infrastructures enable the efficient training of machine/ deep learning models by isolating training data in respective local databases of each collaborative center. Distributed learning can be applied in various forms. In federated learning, each of the collaborators connects to a master server that initializes and updates learning. After initialization, each collaboration center trains a portion of the model on their local data then provides the resulting model weights to the master server. The master server in turn aggregates the weights, updates the model, and shares the updated model weights with the collaborators within the network. Each collaborator then retrains the local models based on the updated weights and sends them back to the master server to close the loop, which operates until a convergence threshold is reached [3,4]. Another form of distributed learning is sequential learning, differing in learning management architecture: 1) learning orchestrated by a cloud server such as the personal health train (PHT; https://www.dtls.nl/fair-data/personal-health-train/) [9,24], or 2) decentralized learning as applied in Chained-Distributed Machines Learning (C-DistriM) [6]. Each iteration in a sequential learning process corresponds to an update of the model from one collaborator. This type of learning is slower when compared to federated learning, where the learning is parallel, but is not subject to the logistical concerns (mainly related to the variation of the internet connection speed across the partners) related to federated learning [25].

In distributed learning data is not visible to the researchers. For this reason, researchers have to rely on the statistical information derived from the local data to build a global model. To reach an optimal performance some modeling steps such as feature selection and inference have to be adapted [4,26,27]. In this regard, the literature has

5

demonstrated both federated learning and distributed learning achieve a comparative performance to traditional centralized learning approaches [4,6,10,28,29].

Machine learning classifiers

In this work we consider three machine learning classifiers, depicted in Figure 5.1, in distributed sequential settings:

1. Support vector machines (SVM), a supervised learning algorithm, applied mostly towards classification, but also for regression and the detection of outliers. SVMs work by establishing two parallel hyperplanes, separating the different classes of the feature space. The best fit is established as the one that maximizes the distance between both hyperplanes [30]. To accommodate data variability (i.e., linearly separable or not) various kernels such as linear, radial basis function have been established to optimize the distance between hyperplanes [31]. In this work we applied linear SVM techniques.

2. Logistic regression is a statistical method used for analyzing a feature space in which there are one or more independent variables that identify a predefined outcome. The assumption is that multiple linear regressions of the independent variables are transformed using a logit function to form a conditional probability of the outcome variable. Logistic regression assumes that the feature space possesses a linear relationship with the outcome, making it a linear algorithm with a nonlinear transform [30].

3. Perceptron, a single-layer neural network used for linear classification. The hidden layer mimics the design of a network of neurons within the human brain. Similarly, a perceptron network predicts classifications based on patterns within a series of input features correlating to a specific outcome[30]. The process is as follow: 1) the input features are multiplied by their corresponding weights, that are randomly selected at the first iteration, 2) sum the results of step (1) to generate a weighted sum, 3) calculate the outcome (output) by applying the weighted sum to the activation function, that maps the outcome into values ranging between two predefined values (labels) such as [0,1].
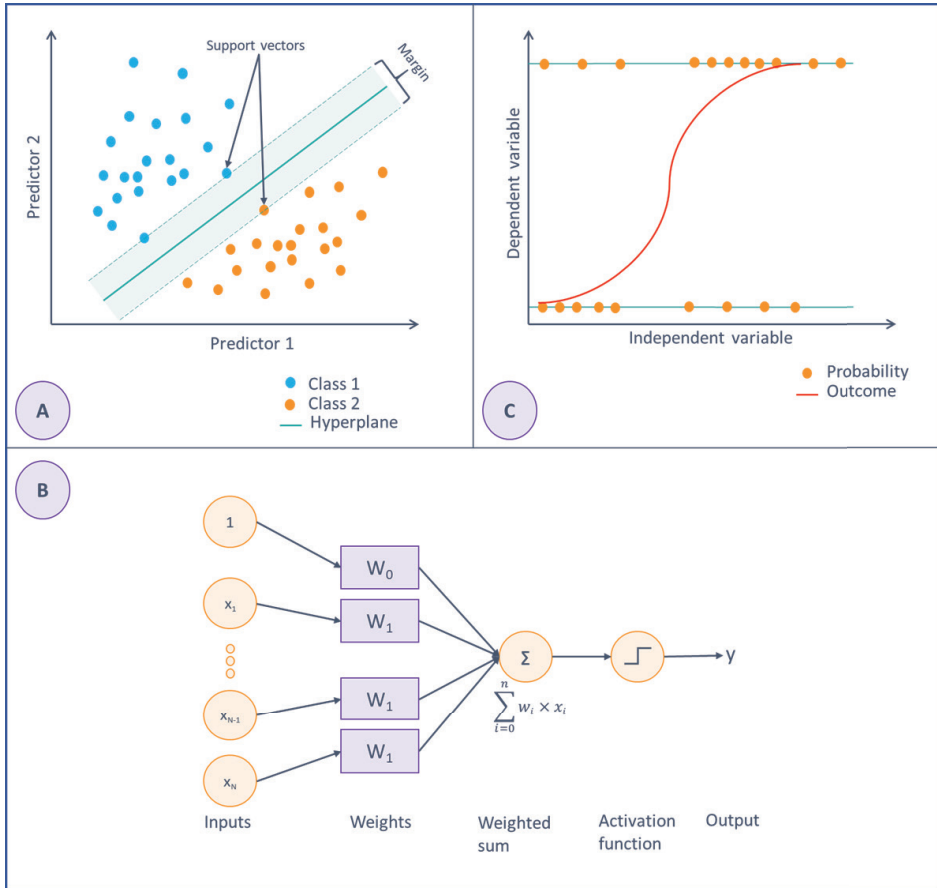
Figure 5.1: A) Support Vector Machine, B) Logistic Regression, C) Perceptron.

Methods

Data

In this study, four open-source datasets were collected from two different public repositories: the UC Irvine machine learning repository (https://archive.ics.uci.edu/ml/index.php) and cancerdata.org (https://www.cancerdata.org/). The characteristics of these datasets are illustrated in Table 5.1. The datasets include:

1. breast cancer Wisconsin dataset [32].

2. Indian liver dataset [33],

3. NSCLC-Radiomics dataset [34,35],

4. Stage III NSCLC dataset [36].

Table 5.1: Dataset characteristics

Dataset	Disease	Outcome	Patients	Training samples	Validation samples	Test samples	Number of Features (after selection)	Feature type
Breast cancer Wisconsin [32]	Breast cancer	Breast mass malignancy	569	341	114	114	15	clinical
Indian liver [33]	Liver disease	Liver disease	583	349	117	117	7	demographics, clinical
NSCLC-Radiomics dataset [34,35]	Non-small-cell lung cancer	2-year survival	421	224	75	75	4	radiomics
Stage III NSCLC [36]	Stage 3 non-small-cell lung cancer	Overall survival (binary)	548	328	110	110	8	clinical, dosimetric

These datasets were used to train and test the selected machine learning classifiers. Each of these datasets consists of a feature space corresponding to a binary outcome, as illustrated in Table 5.1. In addition to these four data sets, we extended our analysis to test sequential distributed learning on deep neural networks applied to smaller sets of (MNIST) dataset [37].

The breast cancer Wisconsin dataset consists of features calculated from a digitized image of a fine needle aspirate (FNA) of a breast mass [38], and an outcome defined as "malignant" or "benign". ANOVA test was used to perform select the robust features. 50% of the features were discarded based on ANOVA's F-ratio, reducing the total number of features from 30 to 15.

The Indian liver dataset [33] consists of a set of demographics and clinical features (all patient records were collected from North East of Andhra Pradesh, India) for patients with liver disease. A Pearson pairwise feature correlation was performed. Highly correlated features (i.e., with Pearson correlation coefficient > 0.7) were discarded, reducing the total number of features from 11 to 8. Four patients had missing values corresponding to one feature, the missing data were imputed based on the mean value of the corresponding feature vector.

The NSCLC-Radiomics dataset [34,35] consists of radiomics features extracted, using RadiomiX (Radiomics/Oncoradiomics SA, Liège, Belgium) based on quantitative image analysis technology, from gross tumor volumes (GTV) of standard CT images corresponding to 422 patients. Gross tumor volume segmentations were performed by trained oncologists. Of the 421 records, 44 subjects were discarded during the radiomic features calculation phase. The discarded subjects had GTV segmentations with multiple unconnected volumes. In these cases, signature feature "compactness" cannot be calculated since it is defined for a single volumetric object. The outcome (survival) in NSCLC-Radiomics was converted into two-year survival (binary). New feature selection was not performed, the four predictive features reported in the original study [39] were used.

Data from [36] is referred to as The Stage III NSCLC dataset. The dataset consists of a combination of clinical, dosimetric features and clinical outcome (survival) for lung cancer patients. Missing data were imputed, using the scikit-learn (version 0.22) imputation transformer. The imputation was based on the mean values of each feature. No feature selection was performed for this dataset, instead predictive features reported in the original study [36] were used to train the models.

As a means to mitigate classifier scaling bias, features in all training sets were independently normalized to the interval [0,1] and the same normalization factor was applied to their respective validation and test sets. The primary objective of this work was to assess model performance variability across unique training scenarios in centralized

vs. distributed SGD training approaches. Improving the prediction performance of for models trained with these datasets was out of the scope of this work.

Experiment design

Three commonly used machine learning (ML) classifiers were selected to conduct this study (Support Vector Machine (SVM), Logistic Regression, and Perceptron). Each classifier satisfies the inclusion criteria:

1. the classifier can be trained in a sequential manner,

2. the classifier has previously been applied and accepted in medical image analysis scientific community [8,40].

The open-source SGDClassifier package (scikit-learn v0.22, Google Summer of Code) in Python (v3.6) was used to implement the selected classifiers [41].

Each dataset was split into training, validation, and test sets (60% training, 20% validation, and 20% testing). Training, validation, and test sets were stratified based on the outcome label to guarantee equal percentage of positive and negative samples on each subset. The validation data was used for hyperparameter tuning and the test set was used evaluate the model performance.

For each dataset, we simulated four training cases:

Centralized: a centralized learning approach where the entirety of the training set was used by a single partner to fit the model – used as the reference for distributed learning approaches.

Case 1: a distributed learning approach composed of 2 partners (2 subsets), randomly distributed between each partner (67% and 33% of the dataset).

Case 2: a distributed learning approach representing an extreme case where each partner contributes with a single datapoint (i.e., from a single patient). In this case the model was updated at each iteration incorporating one additional datapoint.

Case 3: a repeat of Case 2, with the exception of randomly shuffling the dataset to observe the effect the order of the training data (of medical centers) has on the resulting model.

Optimization of Training Parameters

Centralized model

For each classifier and dataset pair, we trained a central model and used it as the reference to compare performance of each corresponding distributed model. Hyperparameters tuning was performed for optimal performance. The primary tuned parameter specified the regularization parameter used to calculate learning rate, herein referred to as alpha (α), and number of iterations (epochs). Default values with respect to the classifier were used for remaining parameters such as tolerance, and penalty. To tune the hyperparameters we defined a set of alpha values ranging between 1e-7 and 1, as illustrated in Figure 5.2. For each classifier:

1. The validation set performance (determined by the area under the curve (AUC) of the Receiver Operating Characteristic curve (ROC)), was estimated for each parameter α.

2. The resulting models were compared based on their performances.

3. The best α parameter was then selected according to the model comparison outcome.

4. Each classifier was subsequently retrained using the best corresponding α parameter.

5. Finally, the performance of the global model was then evaluated against the appropriate test set.

5

Distributed learning models

Hyperparameters tuning was also performed for distributed learning cases, as illustrated in Figure 5.3. Model optimization was performed over 5 key steps:

1. For each simulated partner, estimate the validation set performance (AUC) corresponding to each parameter α.

2. Compare the resulting models based on their performances.

3. Select the best α parameter according to the model comparison outcome.

4. Retrain each classifier using the best corresponding α parameter in a sequential manner.

5. Finally, the performance of the global model was then evaluated against the appropriate test set.
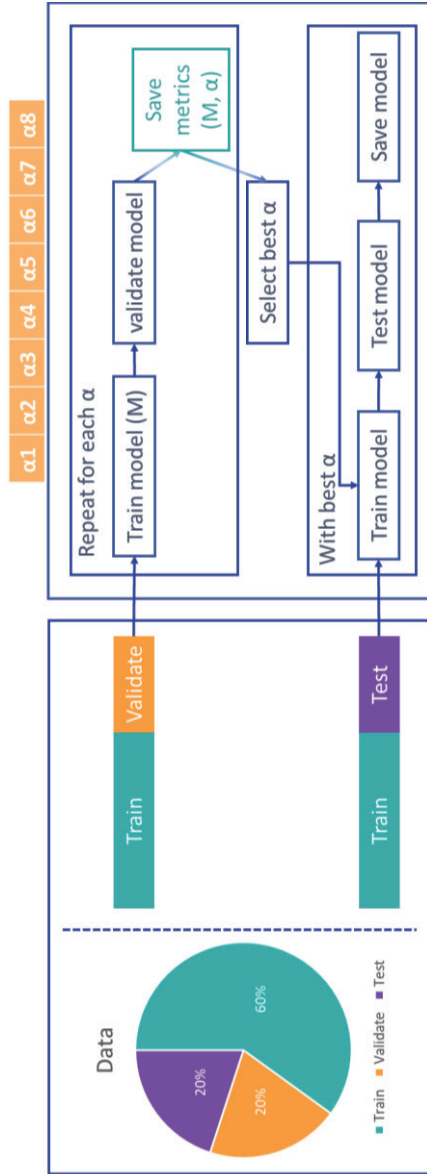
Figure 5.2: Model training design.
1) each dataset is divided into training, validation, and testing sets; **2)** train a model and validate it for each α; **3)** compare the AUCs corresponding to each α; **4)** select the value of α that returned the best AUC; **5)** train the main model using the selected α, **6)** test the model using the test data, **7)** save the model.
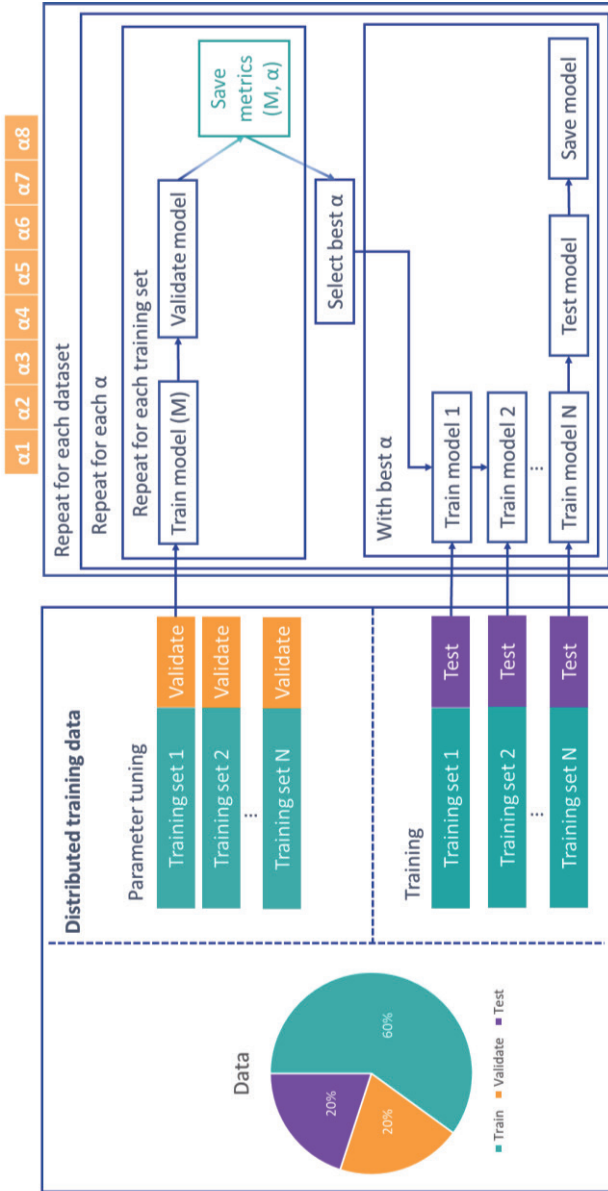
Figure 5.3: Distributed model training design.

1) each dataset is divided into training, validation, and testing sets; **2)** each training set is split into distributed sets; **3)** train a model for each distributed dataset and validate it for each α; **4)** compare the AUCs corresponding to each α; **5)** select the value of α that returned the best AUC; **6)** train the main model using in a sequential manner across the distributed datasets using the selected α, **7)** evaluate all the distributed models using the same test, **8)** save the final model in the queue as the main model.

Finally a pairwise comparison of the final aggregated models AUC values corresponding to each classifier and dataset was performed using DeLong tests [42].

To consider the impact of shuffling the local training datasets, and their size on model performance in cases where partners have multiple datapoints each, we extended the experiments conducted in this study. To sufficiently realize these experiments, we sourced the modified national institute of standards and technology (MNIST) dataset [37], a commonly used large dataset suited to test deep neural networks. Data description, the different data splits, model architecture, and results are available in Supplementary Materials (Section A).

Results

Results based on dataset

The combination of 4 datasets, training cases and 3 model architectures resulted in 48 uniquely trained models. Table 5.2 summaries model performance for each architecture and training use case reported as the AUC and a 95% confidence interval (CI). Models trained with the breast cancer dataset outperformed models trained with other datasets in all model architectures. The Indian dataset had notably better performance in specific training use cases and model architectures when compared to classification performance for either NSCLC dataset. Logistic regression and perceptron architectures had improved performance over SVM for classification in either NSCLC dataset.

Results based on training use case

Figure 5.4 depicts the AUC values corresponding to each study case for each pair of classifier and dataset. For each classifier, the derived AUC values per use case (centralized, case 1, case 2, case 3) trended with a high degree of similarity but were not identical. Shuffling and local dataset size variability produced observable differences in the ROC curves. However, the Pairwise DeLong tests [42] were used to compare the ROC curves for each of the training scenarios and found no statistically significant differences (p-values > 0.05), as summarized in Table 5.3 organized by classifier and training dataset. Furthermore, each model trained in a distributed fashion did not differ significantly from the reference centralized trained model (p-values > 0.05). These results were validated on a CNN classification model using the MNIST dataset. Detailed results of the MNIST experiments are presented in Supplementary Materials (Section A). ROC curves corresponding to each training scenario and dataset is reported in the Supplementary Materials (Section B).

Table 5.2: Discrimination performance (AUC) obtained by training centralized and distributed classifiers (SVM, logistic regression, and Perceptron) using four different datasets (Breast cancer, Indian Liver, NSCLC-Radiomics dataset, and Stage III NSCLC)

Classifier	Training scenario	AUC (95% CI)			
		Breast cancer	Indian Liver	NSCLC-Radiomics dataset	Stage III NSCLC
Support vector machine	Centralized	0.99 (0.98-1)	0.76 (0.68-0.85)	0.64 (0.51-0.77)	0.64 (0.48-0.79)
	Case 1	0.99 (0.99-1)	0.77 (0.69-0.86)	0.64 (0.51-0.77)	0.61 (0.46-0.75)
	Case 2	0.98 (0.98-1)	0.74 (0.65-0.83)	0.65 (0.52-0.77)	0.60 (0.46-0.76)
	Case 3	0.98 (0.97-1)	0.75 (0.67-0.84)	0.62 (0.49-0.75)	0.61 (0.46-0.76)
Logistic Regression	Centralized	0.98 (0.98-1)	0.76 (0.67-0.84)	0.72 (0.61-0.84)	0.70 (0.57-0.82)
	Case 1	0.97 (0.95-0.99)	0.76 (0.67-0.85)	0.71 (0.59-0.82)	0.69 (0.56-0.82)
	Case 2	0.97 (0.94-0.99)	0.73 (0.64-0.82)	0.72 (0.61-0.84)	0.65 (0.52-0.78)
	Case 3	0.99 (0.98-1)	0.74 (0.65-0.83)	0.70 (0.58-0.82)	0.67 (0.55-0.79)
Perceptron	Centralized	0.99 (0.98-1)	0.78 (0.70-0.86)	0.72 (0.61-0.84)	0.70 (0.55-0.84)
	Case 1	0.98 (0.96-1)	0.76 (0.68-0.85)	0.68 (0.55-0.80)	0.69 (0.56-0.82)
	Case 2	0.99 (0.98-1)	0.74 (0.65-0.83)	0.67 (0.54-0.79)	0.67 (0.53-0.81)
	Case 3	0.99 (0.98-1)	0.78 (0.70-0.86)	0.66 (0.54-0.79)	0.69 (0.56-0.81)

5

Results based on classifier architecture

In most cases the average absolute difference in the AUC values were blow 5%. The average difference in the AUC of the centralized training over the Breast cancer, Indian Liver, NSCLC-Radiomics dataset, Stage III NSCLC datasets was reported as 0.67%, 1.75%, 8.33%, and 6.24%, respectively. Differences in the AUC for each training scenario versus each classifier has been summarized in Table S5.1, Supplementary Materials (Section C). The maximum average difference of the AUC values for the distributed learning classifiers per dataset increases up to 8.74%, 7.66%, and 8.64% for case, case 2, case 3, respectively. It should be noted that in extreme cases certain scenarios had AUC differences above 10%, highlighted in Table S5.1 of the Supplementary Materials (Section C). These results suggest the optimal classifier chosen is highly dependent on the characteristics of the dataset.
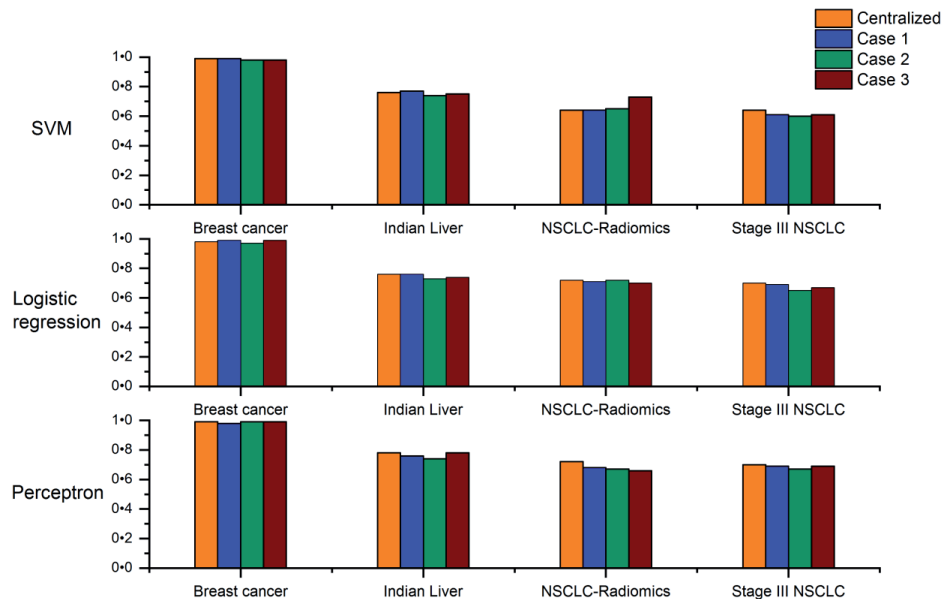
Figure 5.4: AUC of each classifier and dataset pair, for each dataset four different models have been trained.
Centralized, case 1: two partners, case 2: each center holds one patient, case 3: each center holds one patient with a shuffle in the order of the partners while training.

Discussion and future work

High quality datasets with sufficient training datasets are required for machine learning models to converge and generalize [2]. When working with patient data, there are important ethical and legal considerations to be managed, when considering sharing patient data between institutions.

The results presented in this work demonstrate that sequential distributed learning on small, isolated datasets (including extreme cases of model updated using a single datapoint at a time) achieves equivalent performance to models trained in conventional centralized learning. Similar conclusions were observed in the case of multiclass classification using the MNIST dataset [37]. We observed, by applying a pairwise DeLong [42] comparison, that the AUC for distributed learning models do not differ with statistical significance from models trained in centralized scenarios.

The results in Table 5.2 and 5.3 and the ROC curves indicate that there is a difference in the performance of different classifiers, and this difference can vary from one dataset to another. We noted that the average AUC difference between the classifiers can increase up to 8.33%, 8.74%, 7.66%, and 7.66% with respect to each use case (centralized, case 1, case 2, case 3) and dataset (Breast cancer, Indian Liver, NSCLC-Radiomics dataset, Stage

Table 5.3: P-values corresponding to the pairwise Delong test

Model	Test	Breast cancer p-value	Indian Liver p-value	NSCLC-Radiomics dataset p-value	Stage III NSCLC p-value
Support vector machine	Central model vs case 1	0.155	0.594	0.926	0.810
	Central model vs case 2	0.785	0.475	0.871	0.718
	Central model vs case 3	0.082	0.861	0.53	0.760
	Case 1 vs case 2	0.143	0.429	0.969	0.953
	Case 1 vs case 3	0.071	0.710	0.685	0.980
	Case 2 vs case 3	0.271	0.580	0.234	0.941
Logistic Regression	Central model vs case 1	0.061	1	0.422	0.606
	Central model vs case 2	0.076	0.250	0.196	0.541
	Central model vs case 3	0.904	0.538	0.373	0.663
	Case 1 vs case 2	0.425	0.224	0.401	0.652
	Case 1 vs case 3	0.052	0.652	0.823	0.787
	Case 2 vs case 3	0.066	0.677	0.285	0.762
Perceptron	Central model vs case 1	0.147	0.448	0.062	0.600
	Central model vs case 2	0.427	0.267	0.332	0.809
	Central model vs case 3	0.370	0.858	0.210	0.662
	Case 1 vs case 2	0.322	0.274	0.838	0.868
	Case 1 vs case 3	0.329	0.212	0.719	0.936
	Case 2 vs case 3	0.946	0.129	0.848	0.856

5

III NSCLC). Even though this margin may be perceived as inconsequential, the clinical risk of decisions based on predictions must be considered as with all changes in model performance. Cases with AUC differences above our 10% threshold (highlighted in red), indicate that this specific classifier is suboptimal for the dataset in question. Thus, with respect to learning [43], we recommend to select the classifier based on comparative performance of the different centralized and distributed classifiers, or base the selection justified criteria related to the data characteristics that will be used to fit the model.

Previous reports on distributed ensemble learning [11,12], showed the potential of application of this particular type of distributed learning to small siloed datasets. For example, Tuladhar et al. [12] reports that grouping models learned locally from either artificial neural network, SVM, or random forest could efficiently exploit small sets of data to build global models. These results suggest that the application of ensemble learning on small dataset is feasible. While other authors have demonstrated that grouping local logistic regression models, is promising in the case of small datasets [11]. In addition to that, they proposed a model update based on the distributed sets of data information to improve the global model performance on small datasets. Results

of these studies [11,12] showed an overall improvement in global model performance compared to models trained in a single institution data. These results, however, cannot be extended to the case of distributed sequential learning.

Our results demonstrate that sequential distributed learning can be beneficial for the application of AI for outcome prediction in favor of medical institutes holding very small datasets. Practical examples of small datasets can be 1) pediatric cases that tend to suffer from small sample sizes [44], 2) early phase clinical trials where the sample size tend to have around 20 subjects [45], and 3) rare diseases as they have a very low prevalence (< 5/10000 in the European population) [46], making it nearly impossible for a single medical center to collect enough data to train machine learning models. Even with these limitations, and with considerably small datasets (20 to 100 datapoints), researchers have been using machine learning to build diagnosis and prognosis models for rare diseases [47]. The generalizability of trained models is directly related to the quality and quantity of the training data [48]. In this regard, distributed learning provides opportunity to develop generalizable models with small high-quality datasets in multi-center applications while also mitigating the need to share data and maintaining the privacy of all patient information, such as imaging, genomic, or clinical insight.

Batch size is well known to have an effect on final model performance [49], where evidence suggests that large batch size does not always relate to better model performances [50]. Conversely, in distributed learning applications, 1) a smaller batch size has been linked to the privacy of the training data, as it considerably reduces the ability to reproduce training data from shared model weights in case of weights leakage [51], 2) it has been well documented that the order of training partners in a distributed network influences the performance of the model [28]. Our results suggest that the centralized and distributed models are not statistically different. Therefore, we see distributed sequential learning as a viable tool for multicentric precision medicine studies, particularly in applications with small datasets such as rare diseases and could also be applied in pediatrics and early phase clinical trials.

The tuning of each classifier prior training of the final model is an essential step in achieving robust and generalizable models as this is dependent on the nature of data used in training. The need for tuning hyperparameters stems from the fact that the classifiers investigated in this work are using SGD as an optimizer, and thus cannot avoid this optimization step. The main parameter that needs to be optimized is the learning rate; as it controls the manner in which model is modified according to the estimated error at every iteration/update of the model weights. The process of learning rate selection is challenging as a small value theoretically facilitates better performances but in contrast can increase training phase time significantly. On the other hand, a larger learning rate value can result in an unstable training phase, as the model updates very quickly in each iteration causing it to converge to a flat (i.e., less optimal) minima.

Tuning model hyperparameters is also imperative for distributed classifiers, as we showed in this study. Furthermore, we observed that there is a need to investigate different combinations of hyperparameters and number of iterations. Hyperparameter and training settings such as number of iterations, coupled with a set of model selection criteria (based for example on a comparison of model accuracies or model parameters) [52], can be beneficial to reduce the risk of overfitting. However, this leads in turn to one limitation, related to the longer execution time in comparison to traditional centralized training. This increase in time is accounted by the need to investigate all the training data across all the participating partners to set optimal hyperparameters. In addition to this, it is important to consider communication costs, as all parameter tuning and model updates occur over the internet. Characterizing the time required for training is challenging as the duration is highly dependent on each partner's internet bandwidth. Future directions of this work will include analysis to characterize the model training duration in a distributed fashion, identify the scalability of the infrastructure to accommodate larger loads by increasing available computational power either though scale up (additional hardware) or scale out (additional nodes) and investigate the elasticity or ability to dynamically handle varying loads of data.

Conclusion

This study demonstrates 1) the proof-of-concept of sequential distributed learning applied on small sizes of data, narrowed down to a single datapoint at a time 2) the opportunities associated with this type of distributed infrastructures on the application of AI in low prevalence diseases. We simulated three different distributed learning cases using three classifiers and four different datasets. Our results indicate that sequentially training the models using (extremely) small datasets delivers statistically similar performance (p-values > 0.05) in comparison to the conventional centralized approach. This work provides a validation of the potential of distributed learning in case of small datasets and a new opportunity to data driven outcome modeling in rare disease research. Furthermore, this work can be used to continuously update predictive models as new data is available. Finally, future work is planned to estimate and optimize the scalability of sequential distributed learning infrastructures in real world settings.

5

Conflict of interest statement

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Fadila Zerka, Akshayaa Vaidyanathan, Fabio Bottari, Martin Gueuning, Hanif Gabrani-Juma, Mariaelena Occhipinti are salaried employees/receive renumeration from Radiomics (Oncoradiomics SA). Dr Philippe Lambin reports, within and outside the submitted work, grants/sponsored research agreements from Varian medical, Radiomics (Oncoradiomics SA) , ptTheragnostic/DNAmito, Health Innovation Ventures. He received an advisor/presenter fee and/or reimbursement of travel costs/external grant writing fee and/or in kind manpower contribution from Radiomics (Oncoradiomics SA), BHV, Merck, Varian, Elekta, ptTheragnostic and Convert pharmaceuticals. Dr Lambin has shares in the company Radiomics (Oncoradiomics SA), Convert pharmaceuticals SA and The Medical Cloud Company SPRL and is co-inventor of two issued patents with royalties on radiomics (PCT/NL2014/050248, PCT/NL2014/050728) licensed to Radiomics (Oncoradiomics SA) and one issue patent on mtDNA (PCT/EP2014/059089) licensed to ptTheragnostic/DNAmito, three non-patented invention (softwares) licensed to ptTheragnostic/DNAmito, Radiomics (Oncoradiomics SA) and Health Innovation Ventures and three non-issues, non licensed patents on Deep Learning-Radiomics and LSRT (N2024482, N2024889, N2024889. Ralph T.H. Leijenaar has shares in the company Radiomics (Oncoradiomics SA) and is co-inventor of an issued patent with royalties on radiomics (PCT/NL2014/050728) licensed to Radiomics (Oncoradiomics SA). Sean Walsh and Wim Vos have shares in the company Radiomics (Oncoradiomics SA). Michel Dumontier has shares in The Medical Cloud Company SPRL. Rest of the co-authors have no known competing financial interests or personal relationships to declare.

References

[1]   P. Lambin, R.G.P.M. van Stiphout, M.H.W. Starmans, E. Rios-Velazquez, G. Nalbantov, H.J.W.L. Aerts, E. Roelofs, W. van Elmpt, P.C. Boutros, P. Granone, V. Valentini, A.C. Begg, D. De Ruysscher, A. Dekker, Predicting outcomes in radiation oncology—multifactorial decision support systems, Nature Reviews Clinical Oncology. 10 (2013) 27–40. https://doi.org/10.1038/nrclinonc.2012.196.

[2]   P. Lambin, E. Roelofs, B. Reymen, E.R. Velazquez, J. Buijsen, C.M.L. Zegers, S. Carvalho, R.T.H. Leijenaar, G. Nalbantov, C. Oberije, M. Scott Marshall, F. Hoebers, E.G.C. Troost, R.G.P.M. van Stiphout, W. van Elmpt, T. van der Weijden, L. Boersma, V. Valentini, A. Dekker, 'Rapid Learning health care in oncology' – An approach towards decision support systems enabling customised radiotherapy', Radiotherapy and Oncology. 109 (2013) 159–164. https://doi.org/10.1016/j.radonc.2013.07.007.

[3]   T.M. Deist, A. Jochems, J. van Soest, G. Nalbantov, C. Oberije, S. Walsh, M. Eble, P. Bulens, P. Coucke, W. Dries, A. Dekker, P. Lambin, Infrastructure and distributed learning methodology for privacy-preserving multi-centric rapid learning health care: euroCAT, Clinical and Translational Radiation Oncology. 4 (2017) 24–31. https://doi.org/10.1016/j.ctro.2016.12.004.

[4]   M. Bogowicz, A. Jochems, T.M. Deist, S. Tanadini-Lang, S.H. Huang, B. Chan, J.N. Waldron, S. Bratman, B. O'Sullivan, O. Riesterer, G. Studer, J. Unkelbach, S. Barakat, R.H. Brakenhoff, I. Nauta, S.E. Gazzani, G. Calareso, K. Scheckenbach, F. Hoebers, F.W.R. Wesseling, S. Keek, S. Sanduleanu, R.T.H. Leijenaar, M.R. Vergeer, C.R. Leemans, C.H.J. Terhaard, M.W.M. van den Brekel, O. Hamming-Vrieze, M.A. van der Heijden, H.M. Elhalawani, C.D. Fuller, M. Guckenberger, P. Lambin, Privacy-preserving distributed learning of radiomics to predict overall survival and HPV status in head and neck cancer, Scientific Reports. 10 (2020) 4542. https://doi.org/10.1038/s41598-020-61297-4.

[5]   M. Kirienko, M. Sollini, G. Ninatti, D. Loiacono, E. Giacomello, N. Gozzi, F. Amigoni, L. Mainardi, P.L. Lanzi, A. Chiti, Distributed learning: a reliable privacy-preserving strategy to change multicenter collaborations using AI, Eur J Nucl Med Mol Imaging. (2021). https://doi.org/10.1007/s00259-021-05339-7.

[6]   F. Zerka, V. Urovi, A. Vaidyanathan, S. Barakat, R.T.H. Leijenaar, S. Walsh, H. Gabrani-Juma, B. Miraglio, H.C. Woodruff, M. Dumontier, P. Lambin, Blockchain for Privacy Preserving and Trustworthy Distributed Machine Learning in Multicentric Medical Imaging (C-DistriM), IEEE Access. 8 (2020) 183939–183951. https://doi.org/10.1109/ACCESS.2020.3029445.

[7]   S. Lugan, P. Desbordes, L.X.R. Tormo, A. Legay, B. Macq, Secure Architectures Implementing Trusted Coalitions for Blockchained Distributed Learning (TCLearn), ArXiv:1906.07690 [Cs, Stat]. (2019). http://arxiv.org/abs/1906.07690 (accessed August 30, 2019).

[8]   F. Zerka, S. Barakat, S. Walsh, M. Bogowicz, R.T.H. Leijenaar, A. Jochems, B. Miraglio, D. Townend, P. Lambin, Systematic Review of Privacy-Preserving Distributed Machine Learning From Federated Databases in Health Care, JCO Clinical Cancer Informatics. (2020) 184–200. https://doi.org/10.1200/CCI.19.00047.

[9]   T.M. Deist, F.J.W.M. Dankers, P. Ojha, M. Scott Marshall, T. Janssen, C. Faivre-Finn, C. Masciocchi, V. Valentini, J. Wang, J. Chen, Z. Zhang, E. Spezi, M. Button, J. Jan Nuyttens, R. Vernhout, J. van Soest, A. Jochems, R. Monshouwer, J. Bussink, G. Price, P. Lambin, A. Dekker, Distributed learning on 20 000+ lung cancer patients – The Personal Health Train, Radiotherapy and Oncology. 144 (2020) 189–200. https://doi.org/10.1016/j.radonc.2019.11.019.

[10]   A. Jochems, T.M. Deist, I. El Naqa, M. Kessler, C. Mayo, J. Reeves, S. Jolly, M. Matuszak, R. Ten Haken, J. van Soest, C. Oberije, C. Faivre-Finn, G. Price, D. de Ruysscher, P. Lambin, A. Dekker, Developing and Validating a Survival Prediction Model for NSCLC Patients Through Distributed Learning Across 3 Countries, International Journal of Radiation Oncology*Biology*Physics. 99 (2017) 344–352. https://doi.org/10.1016/j.ijrobp.2017.04.021.

[11]   T.-T. Kuo, J. Kim, R.A. Gabriel, Privacy-preserving model learning on a blockchain network-of-networks, Journal of the American Medical Informatics Association. 27 (2020) 343–354. https://doi.org/10.1093/jamia/ocz214.

[12]   A. Tuladhar, S. Gill, Z. Ismail, N.D. Forkert, Building machine learning models without sharing patient data: A simulation-based analysis of distributed learning by ensembling, Journal of Biomedical Informatics. 106 (2020) 103424. https://doi.org/10.1016/j.jbi.2020.103424.

[13]   N. Ketkar, Stochastic Gradient Descent, in: N. Ketkar (Ed.), Deep Learning with Python: A Hands-on Introduction, Apress, Berkeley, CA, 2017: pp. 113–132. https://doi.org/10.1007/978-1-4842-2766-4_8.

[14]   J. Weese, C. Lorenz, Four challenges in medical image analysis from an industrial perspective, Medical Image Analysis. 33 (2016) 44–49. https://doi.org/10.1016/j.media.2016.06.023.

[15]   G. Vegas-Sánchez-Ferrero, M.J. Ledesma-Carbayo, G.R. Washko, R. San José Estépar, Harmonization of chest CT scans for different doses and reconstruction methods, Med. Phys. 46 (2019) 3117–3132. https://doi.org/10.1002/mp.13578.

5

[16]   R. Da-ano, I. Masson, F. Lucia, M. Doré, P. Robin, J. Alfieri, C. Rousseau, A. Mervoyer, C. Reinhold, J. Castelli, R. De Crevoisier, J.F. Rameé, O. Pradier, U. Schick, D. Visvikis, M. Hatt, Performance comparison of modified ComBat for harmonization of radiomic features for multicenter studies, Sci Rep. 10 (2020) 10248. https://doi.org/10.1038/s41598-020-66110-w.

[17]   A.F. Frangi, J.A. Schnabel, C. Davatzikos, C. Alberola-López, G. Fichtinger, eds., Medical Image Computing and Computer Assisted Intervention – MICCAI 2018: 21st International Conference, Granada, Spain, September 16-20, 2018, Proceedings, Part I, Springer International Publishing, Cham, 2018. https://doi.org/10.1007/978-3-030-00928-1.

[18]   R. Banzi, S. Canham, W. Kuchinke, K. Krleza-Jeric, J. Demotes-Mainard, C. Ohmann, Evaluation of repositories for sharing individual-participant data from clinical studies, Trials. 20 (2019) 169. https://doi.org/10.1186/s13063-019-3253-3.

[19]   C. Faviez, X. Chen, N. Garcelon, A. Neuraz, B. Knebelmann, R. Salomon, S. Lyonnet, S. Saunier, A. Burgun, Diagnosis support systems for rare diseases: a scoping review, Orphanet Journal of Rare Diseases. 15 (2020) 94. https://doi.org/10.1186/s13023-020-01374-z.

[20]   E. Turro, W.J. Astle, K. Megy, S. Gräf, D. Greene, O. Shamardina, H.L. Allen, A. Sanchis-Juan, M. Frontini, C. Thys, J. Stephens, R. Mapeta, O.S. Burren, K. Downes, M. Haimel, S. Tuna, S.V.V. Deevi, T.J. Aitman, D.L. Bennett, P. Calleja, K. Carss, M.J. Caulfield, P.F. Chinnery, P.H. Dixon, D.P. Gale, R. James, A. Koziell, M.A. Laffan, A.P. Levine, E.R. Maher, H.S. Markus, J. Morales, N.W. Morrell, A.D. Mumford, E. Ormondroyd, S. Rankin, A. Rendon, S. Richardson, I. Roberts, N.B.A. Roy, M.A. Saleem, K.G.C. Smith, H. Stark, R.Y.Y. Tan, A.C. Themistocleous, A.J. Thrasher, H. Watkins, A.R. Webster, M.R. Wilkins, C. Williamson, J. Whitworth, S. Humphray, D.R. Bentley, N. Kingston, N. Walker, J.R. Bradley, S. Ashford, C.J. Penkett, K. Freson, K.E. Stirrups, F.L. Raymond, W.H. Ouwehand, Whole-genome sequencing of patients with rare diseases in a national health system, Nature. 583 (2020) 96–102. https://doi.org/10.1038/s41586-020-2434-2.

[21]   M. Gymrek, A.L. McGuire, D. Golan, E. Halperin, Y. Erlich, Identifying Personal Genomes by Surname Inference, Science. 339 (2013) 321–324. https://doi.org/10.1126/science.1229566.

[22]   L. Rocher, J.M. Hendrickx, Y.-A. de Montjoye, Estimating the success of re-identifications in incomplete datasets using generative models, Nature Communications. 10 (2019) 3069. https://doi.org/10.1038/s41467-019-10933-3.

[23]   W. Xia, Z. Wan, Z. Yin, J. Gaupp, Y. Liu, E.W. Clayton, M. Kantarcioglu, Y. Vorobeychik, B.A. Malin, It's all in the timing: calibrating temporal penalties for biomedical data sharing, Journal of the American Medical Informatics Association. 25 (2018) 25–31. https://doi.org/10.1093/jamia/ocx101.

[24]   Personal Health Train, Dutch Techcentre for Life Sciences. (n.d.). https://www.dtls.nl/fair-data/personal-health-train/ (accessed January 28, 2021).

[25]   S. Park, Y. Suh, J. Lee, FedPSO: Federated Learning Using Particle Swarm Optimization to Reduce Communication Costs, Sensors (Basel). 21 (2021). https://doi.org/10.3390/s21020600.

[26]   V. Bolón-Canedo, N. Sánchez-Maroño, A. Alonso-Betanzos, Distributed feature selection: An application to microarray data classification, Applied Soft Computing. 30 (2015) 136–150. https://doi.org/10.1016/j.asoc.2015.01.035.

[27]   S.M. Ayyad, A.I. Saleh, L.M. Labib, A new distributed feature selection technique for classifying gene expression data, Int. J. Biomath. 12 (2019) 1950039. https://doi.org/10.1142/S1793524519500396.

[28]   K. Chang, N. Balachandar, C. Lam, D. Yi, J. Brown, A. Beers, B. Rosen, D.L. Rubin, J. Kalpathy-Cramer, Distributed deep learning networks among institutions for medical imaging, Journal of the American Medical Informatics Association. 25 (2018) 945–954. https://doi.org/10.1093/jamia/ocy017.

[29]   A. Choudhury, S. Theophanous, P.-I. Lønne, R. Samuel, M.G. Guren, M. Berbee, P. Brown, J. Lilley, J. van Soest, A. Dekker, A. Gilbert, E. Malinen, L. Wee, A.L. Appelt, Predicting outcomes in anal cancer patients using multi-centre data and distributed learning – A proof-of-concept study, Radiotherapy and Oncology. 159 (2021) 183–189. https://doi.org/10.1016/j.radonc.2021.03.013.

[30]   M. Kuhn, K. Johnson, Applied Predictive Modeling, Springer New York, New York, NY, 2013. https://doi.org/10.1007/978-1-4614-6849-3.

[31]   M. Achirul Nanda, K. Boro Seminar, D. Nandika, A. Maddu, A Comparison Study of Kernel Functions in the Support Vector Machine and Its Application for Termite Detection, Information. 9 (2018) 5. https://doi.org/10.3390/info9010005.

[32]   UCI Machine Learning Repository: Breast Cancer Wisconsin (Diagnostic) Data Set, (n.d.). https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Diagnostic%29 (accessed January 6, 2021).

[33] ILPD (Indian Liver Patient Dataset) - dataset by uci, Data.World. (n.d.). https://data.world/uci/ilpd-indian-liver-patient-dataset (accessed January 6, 2021).

[34] H.J.W.L. Aerts, E. Rios Velazquez, R.T.H. Leijenaar, P. Chintan, P. Grossmann, S. Carvalho, J. Bussink, R. Monshouwer, B. Haibe-Kains, D. Rietveld, F. Hoebers, M. M. Rietbergen, C.R. Leemans, A. Dekker, J. Quackenbush, R. Barkovich, P. Lambin, Data From NSCLC-Radiomics [Data set], (2019). https://doi.org/10.7937/K9/TCIA.2015.PF0M9REI.

[35] K. Clark, B. Vendt, K. Smith, J. Freymann, J. Kirby, P. Koppel, S. Moore, S. Phillips, D. Maffitt, M. Pringle, L. Tarbox, F. Prior, The Cancer Imaging Archive (TCIA): Maintaining and Operating a Public Information Repository, J Digit Imaging. 26 (2013) 1045–1057. https://doi.org/10.1007/s10278-013-9622-7.

[36] C. Oberije, D. De Ruysscher, R. Houben, M. van de Heuvel, W. Uyterlinde, J.O. Deasy, J. Belderbos, A.-M.C. Dingemans, A. Rimner, S. Din, P. Lambin, A Validated Prediction Model for Overall Survival From Stage III Non-Small Cell Lung Cancer: Toward Survival Prediction for Individual Patients, International Journal of Radiation Oncology*Biology*Physics. 92 (2015) 935–944. https://doi.org/10.1016/j.ijrobp.2015.02.048.

[37] Y. Lecun, Gradient-Based Learning Applied to Document Recognition, PROCEEDINGS OF THE IEEE. 86 (1998) 47.

[38] K.P. Bennett, O.L. Mangasarian, Robust linear programming discrimination of two linearly inseparable sets, Optimization Methods and Software. 1 (1992) 23–34. https://doi.org/10.1080/10556789208805504.

[39] H.J.W.L. Aerts, E.R. Velazquez, R.T.H. Leijenaar, C. Parmar, P. Grossmann, S. Carvalho, J. Bussink, R. Monshouwer, B. Haibe-Kains, D. Rietveld, F. Hoebers, M.M. Rietbergen, C.R. Leemans, A. Dekker, J. Quackenbush, R.J. Gillies, P. Lambin, Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach, Nat Commun. 5 (2014) 4006. https://doi.org/10.1038/ncomms5006.

[40] J.A.M. Sidey-Gibbons, C.J. Sidey-Gibbons, Machine learning in medicine: a practical introduction, BMC Medical Research Methodology. 19 (2019) 64. https://doi.org/10.1186/s12874-019-0681-4.

[41] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, É. Duchesnay, Scikit-learn: Machine Learning in Python, Journal of Machine Learning Research. 12 (2011) 2825–2830.

[42] E.R. DeLong, D.M. DeLong, D.L. Clarke-Pearson, Comparing the Areas under Two or More Correlated Receiver Operating Characteristic Curves: A Nonparametric Approach, Biometrics. 44 (1988) 837. https://doi.org/10.2307/2531595.

[43] T.M. Deist, F.J.W.M. Dankers, G. Valdes, R. Wijsman, I.-C. Hsu, C. Oberije, T. Lustberg, J. van Soest, F. Hoebers, A. Jochems, I. El Naqa, L. Wee, O. Morin, D.R. Raleigh, W. Bots, J.H. Kaanders, J. Belderbos, M. Kwint, T. Solberg, R. Monshouwer, J. Bussink, A. Dekker, P. Lambin, Machine learning algorithms for outcome prediction in (chemo)radiotherapy: An empirical comparison of classifiers, Medical Physics. 45 (2018) 3449–3459. https://doi.org/10.1002/mp.12967.

[44] V. Rahimzadeh, C. Schickhardt, B.M. Knoppers, K. Sénécal, D.F. Vears, C.V. Fernandez, S. Pfister, S. Plon, S. Terry, J. Williams, M.S. Williams, M. Cornel, J.M. Friedman, Key Implications of Data Sharing in Pediatric Genomics, JAMA Pediatr. 172 (2018) 476. https://doi.org/10.1001/jamapediatrics.2017.5500.

[45] Phase I Trials - an overview | ScienceDirect Topics, (n.d.). https://www.sciencedirect.com/topics/biochemistry-genetics-and-molecular-biology/phase-i-trials (accessed July 7, 2021).

[46] D. Taruscio, L. Vittozzi, A. Rocchetti, P. Torreri, L. Ferrari, The Occurrence of 275 Rare Diseases and 47 Rare Disease Groups in Italy. Results from the National Registry of Rare Diseases, Int J Environ Res Public Health. 15 (2018). https://doi.org/10.3390/ijerph15071470.

[47] J. Schaefer, M. Lehne, J. Schepers, F. Prasser, S. Thun, The use of machine learning in rare diseases: a scoping review, Orphanet Journal of Rare Diseases. 15 (2020) 145. https://doi.org/10.1186/s13023-020-01424-6.

[48] T. Lustberg, J. van Soest, A. Jochems, T. Deist, Y. van Wijk, S. Walsh, P. Lambin, A. Dekker, Big Data in radiation therapy: challenges and opportunities, The British Journal of Radiology. 90 (2017) 20160689. https://doi.org/10.1259/bjr.20160689.

[49] X. Qian, D. Klabjan, The Impact of the Mini-batch Size on the Variance of Gradients in Stochastic Gradient Descent, ArXiv:2004.13146 [Cs, Math]. (2020). http://arxiv.org/abs/2004.13146 (accessed January 18, 2021).

5

[50]   I. Kandel, M. Castelli, The effect of batch size on the generalizability of the convolutional neural networks on a histopathology dataset, ICT Express. 6 (2020) 312–315. https://doi.org/10.1016/j.icte.2020.04.010.

[51]   M. Abadi, A. Chu, I. Goodfellow, H.B. McMahan, I. Mironov, K. Talwar, L. Zhang, Deep Learning with Differential Privacy, Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security - CCS'16. (2016) 308–318. https://doi.org/10.1145/2976749.2978318.

[52]   M.W. Browne, Cross-Validation Methods | Elsevier Enhanced Reader, (n.d.). https://doi.org/10.1006/jmps.1999.1279.

Supplementary materials

Section A
MNIST dataset

Methods

Data

MNIST, is a publicly available dataset, that was developed to evaluate machine/deep learning models trained to classify handwritten digits [1]. The MNIST dataset consists of 60000 training images and 10000 test images (28x28 pixels) of handwritten digits (0-9). It has been widely used in the development and validation of novel handwritten digit detection models.

Model

The model is based on Convolutional Neural Networks (CNN). CNNs are a an advanced and more complex implementation of simple neural networks. These algorithms are commonly used to classify and analyze different types of imaging depending on the desired outcome.

The model architecture consists of an input layer of shape (28, 28, 1), followed by two convolutional layers, by pooling layer, and a fully connected layer, in addition to an output layer consisting of a SoftMax activation function. Multiple filters are used at each convolutional layer, for different types of feature extraction.

Training

- A centralized model was training using all the 60000 training images and tested using the 10000 test images.

Different distributed learning scenarios were simulated using the training set, to test the effect of training using different dataset sizes, all the models were tested using the same test set.

- **Case 01:** the training data was split into two centers of set 1 (n = 40200) and set 2 (n = 19800) respectively, the model was trained by fitting the first set of data and then update this model using the second set of data.
- **Case 01 shuffle:** in this case the exact same data split used in "Case 01" was employed, however the training order was changed, meaning the model was trained using set 2 (n = 19800) and then updated using set 1 (n = 40200).
- **Case 02:** the training data was split into four centers: set 1 (n = 19800), set 2 (n = 13266), set 3 (n = 18045), and set 4 (n = 8889) respectively. The model

5

was trained by fitting set 1, then updated using set 2, followed by set 3, and finally update the model using the set 4.

- **Case 02 shuffle:** in this case the same data split used in "Case 02" was employed with a random shuffle of the training order. The training odder was set 4 (n = 8889), followed by set 1 (n = 19800), followed by set 3 (n = 18045) and finally the model was updated using set 2 (n = 13266).
- **Case 03:** the training data was split into ten centers: set 1 (n = 6534), set 2 (n = 13266), set 3 (n = 8889), set 4 (n = 5955), set 5 (n = 8100), set 6 (n = 3990), set 7 (n = 4378), set 8 (n = 2934), set 9 (n = 1965), and set 10 (n = 3989). The model was trained by fitting all the training sets in ascending order from set 1 to set 10.
- **Case 03 shuffle:** in this case the same data split used in "Case 03" was employed with a random shuffle of the training order. The new training odder was set 4 (n = 5955), followed by set 2 (n = 13266), set 7 (n = 4378), set 3 (n = 8889), set 6 (n = 3990), set 10 (n = 3989), set 9 (n = 1965), set 8 (n = 2934), set 5 (n = 8100), and finally the model was updated using set 1 (n = 6534).
- **Case 04:** the training data was split into twenty centers: set 1 (n = 13266), set 2 (n = 8889), set 3 (n = 5955), set 4 (n = 3990), set 5 (n = 2673), set 6 (n = 1791), set 7 (n = 1200), set 8 (n = 804), set 9 (n = 539), set 10 (n =732), set 11 (n = 361), set 12 (n = 6534), set 13 (n = 4378), set 14 (n = 2934), set 15 (n = 1965), set 16 (n = 1317), set 17 (n = 882), set 18 (n = 591), set 19 (n = 803), and set 20 (n = 396). The model was trained by fitting all the training sets in ascending order from set 1 to set 20.
- **Case 04 shuffle:** in this case the same data split used in "Case 04" was employed with a random shuffle of the training order. The new training odder was set 13 (n = 4378), set 14 (n = 2934), set 20 (n = 396), set 3 (n = 5955), set 11 (n = 361), set 10 (n =732), set 7 (n = 1200), set 18 (n = 591), set 17 (n = 882), set 1 (n = 13266), set 9 (n = 539), set 2 (n = 8889), set 6 (n = 1791), set 4 (n = 3990), set 12 (n = 6534), set 5 (n = 2673), set 16 (n = 1317), set 19 (n = 803), set 8 (n = 804), and set 15 (n = 1965).

Results

The results of the model training for the centralized model and all the cases described in section (c. Training) of the Appendix are depicted the Figures S5.1, S5.2, S5.3, S5.4, and S5.5 respectively. The results obtained with the MNIST dataset support the results obtained with the four datasets and machines leaning classifiers reported in the main paper.
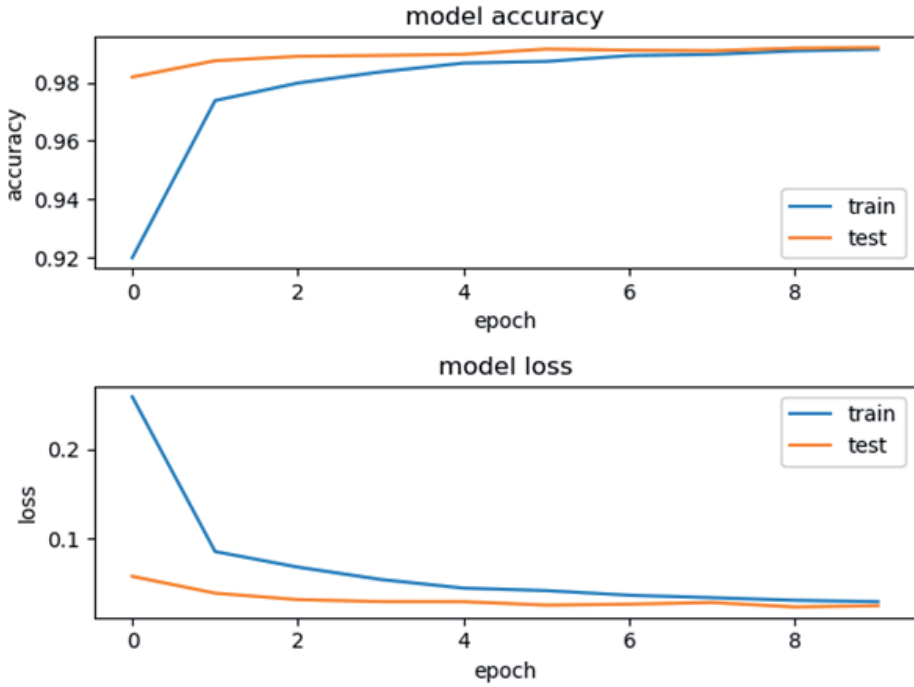
Figure S5.1: Central model: model accuracy and loss on the on the training and test sets.
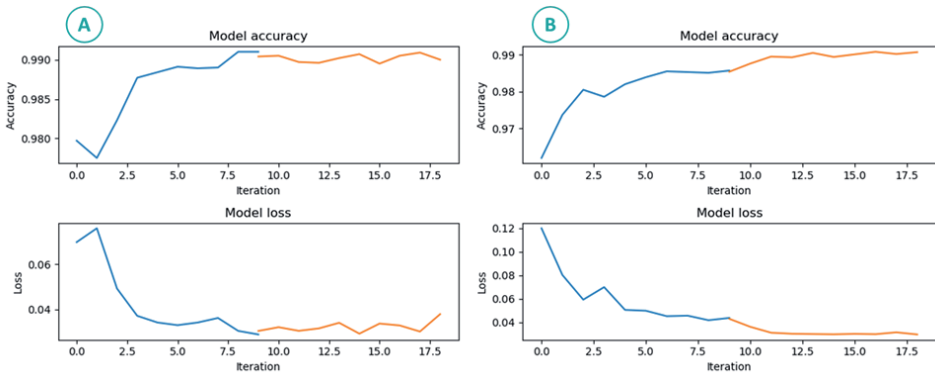

Figure S5.2: A) case 01: model accuracy and loss on the test set, B) case 01 shuffle: model accuracy and loss on the test set.
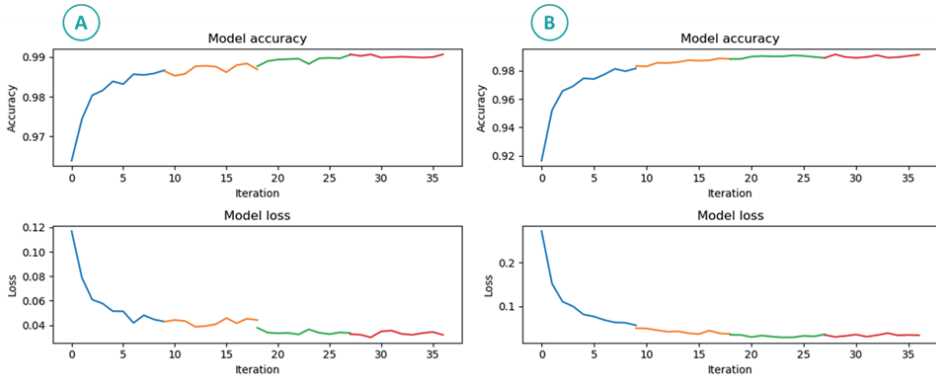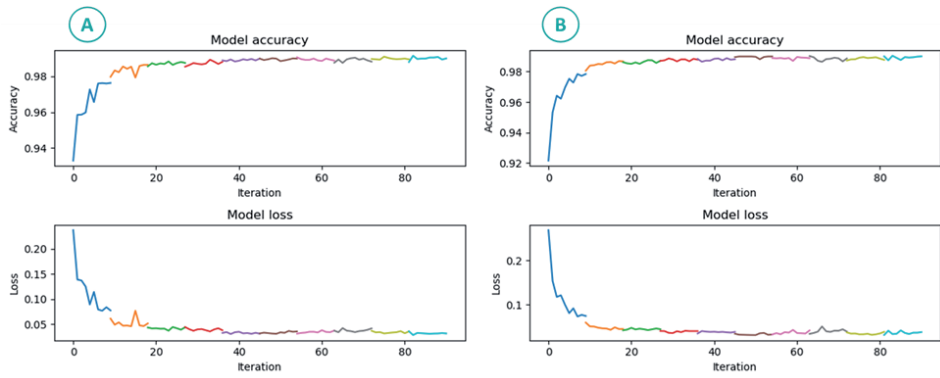
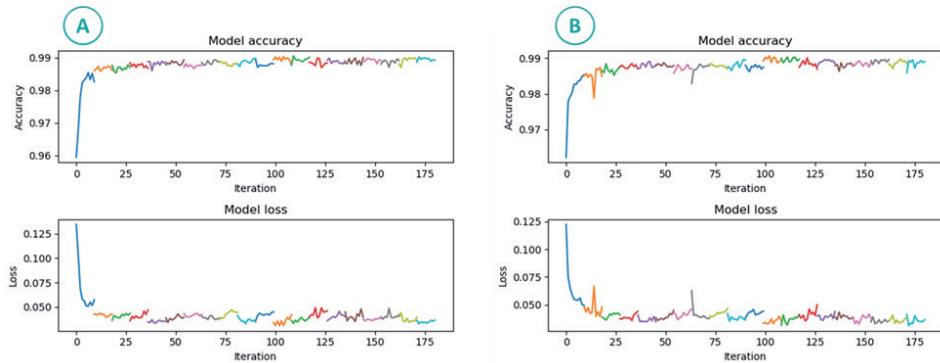**Figure S5.3: A)** case 02: model accuracy and loss on the test set, **B)** case 02 shuffle: model accuracy and loss on the test set.

Figure S5.4: A) case 03: model accuracy and loss on the test set, **A)** case 03 shuffle: model accuracy and loss on the test set.

Figure S5.5: A) case 03: model accuracy and loss on the test set, **B)** case 03 shuffle: model accuracy and loss on the test set.

Section B
ROC curves

BREAST CANCER Wisconsin Dataset
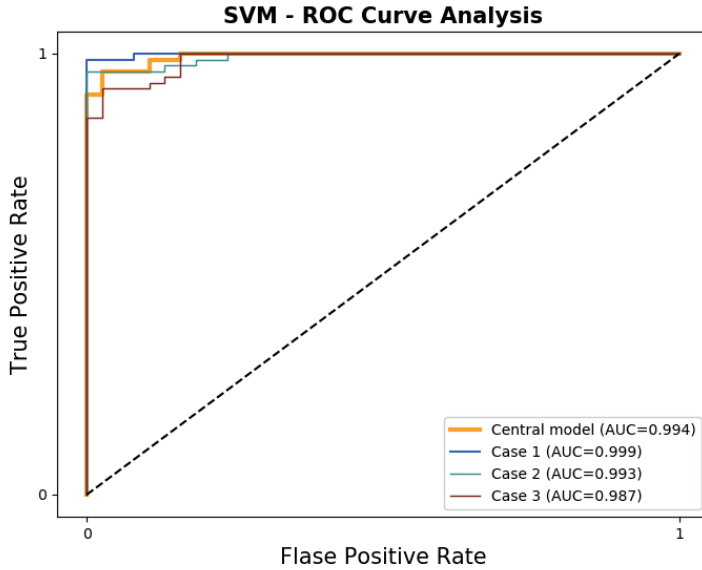
Figure S5.6: Receiver operator characteristic curves for breast mass classification model (SVM) trained using centralized learning and distributed learning.
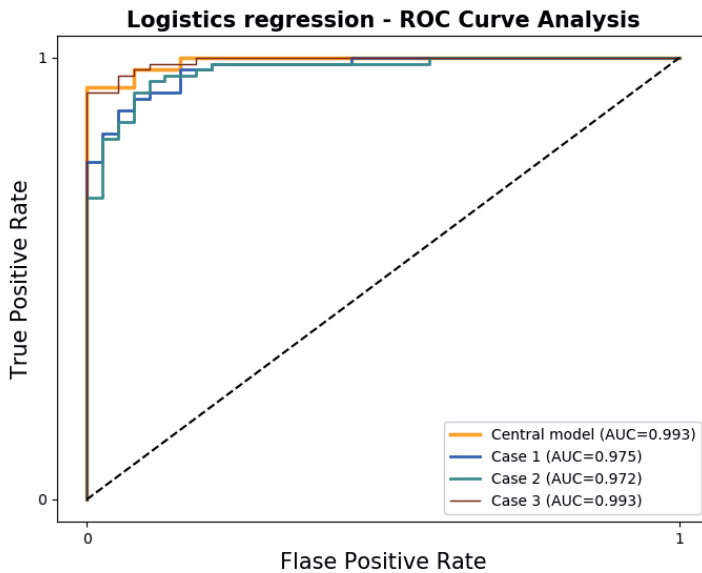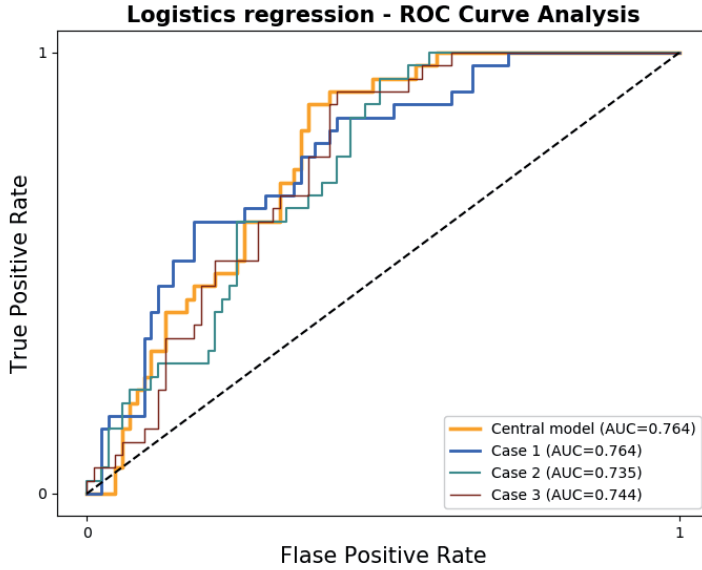

Figure S5.7: Receiver operator characteristic curves for breast mass classification model (Logistic regression) trained using centralized learning and distributed learning.
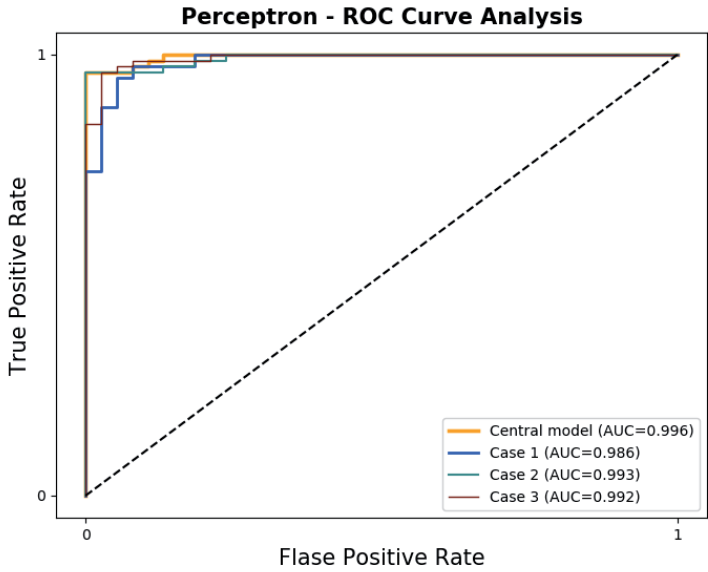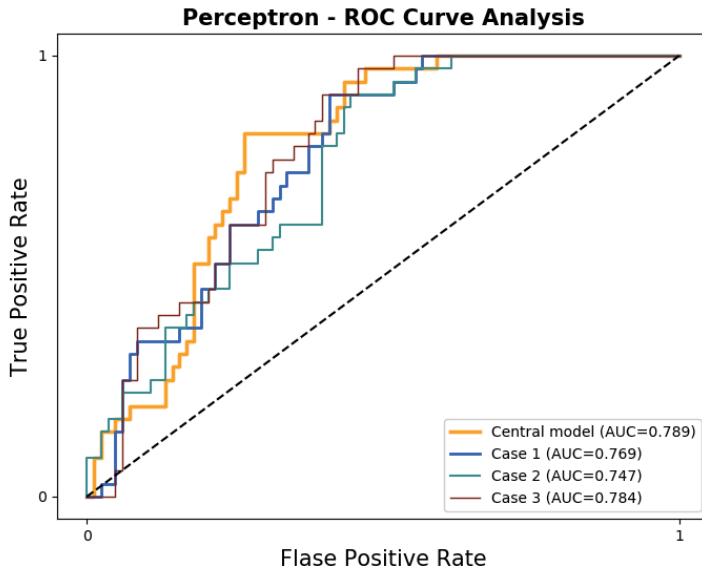
Figure S5.8: Receiver operator characteristic curves for breast mass classification model (Perceptron) trained using centralized learning and distributed learning.
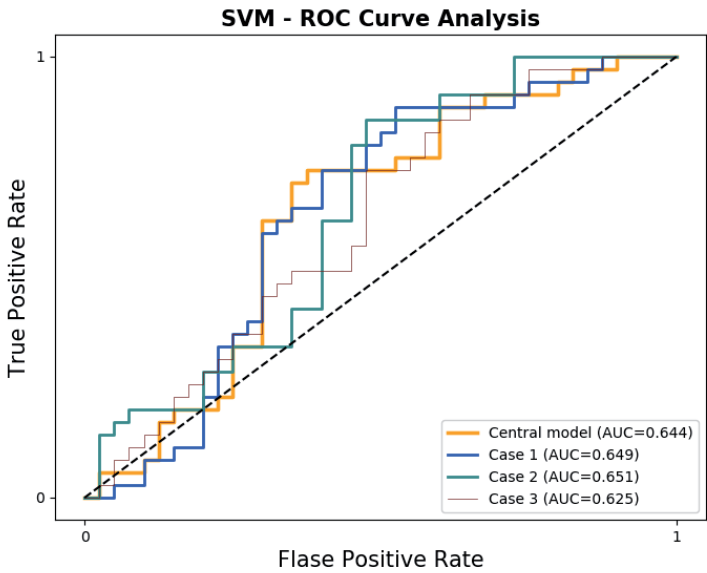
INDIAN LIVER dataset
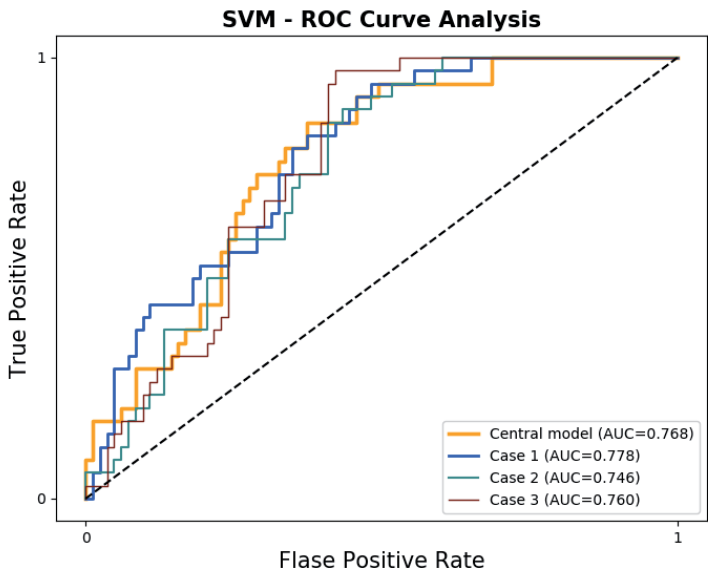
Figure S5.9: Receiver operator characteristic curves for liver disease detection model (SVM) trained using centralized learning and distributed learning.
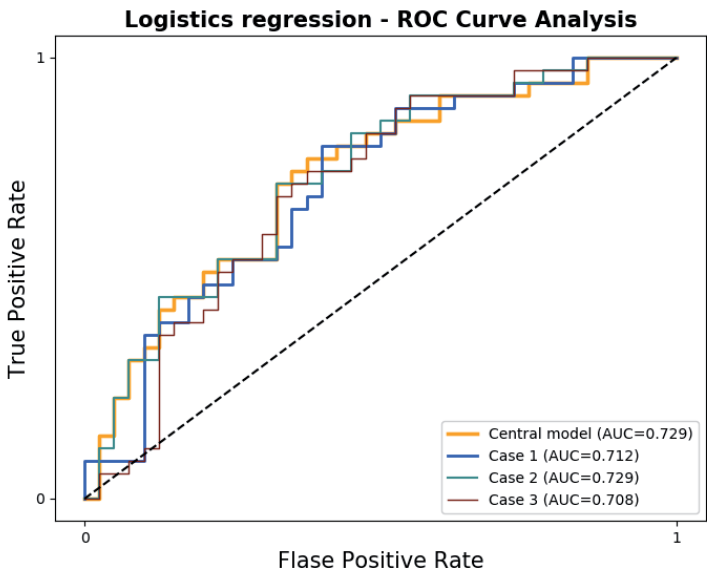
Figure S5.10: Receiver operator characteristic curves for liver disease detection model (Logistic regression) trained using centralized learning and distributed learning.

Figure S5.11: Receiver operator characteristic curves for liver disease detection model (Perceptron) trained using centralized learning and distributed learning.

NSCLC-RADIOMICS dataset

Figure S5.12: Receiver operator characteristic curves for two-year survival model (SVM) trained using centralized learning and distributed learning.

Figure S5.13: Receiver operator characteristic curves for two-year survival model (Logistic regression) trained using centralized learning and distributed learning.
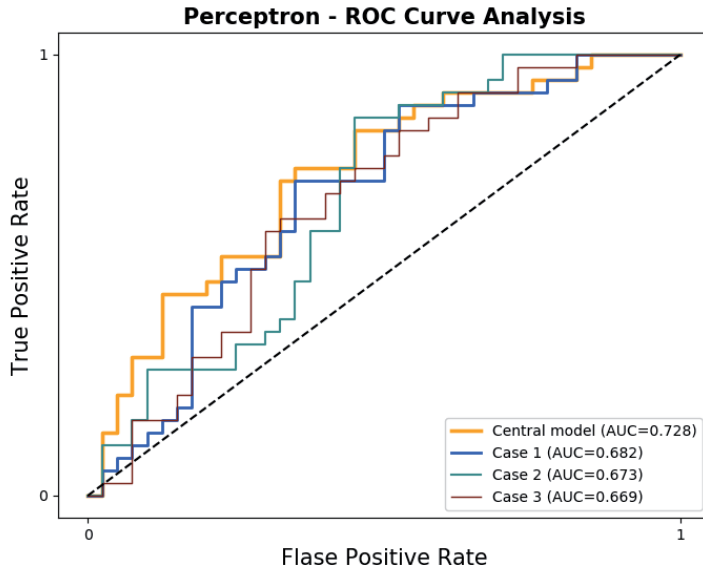
Figure S5.14: Receiver operator characteristic curves for two-year survival model (Perceptron) trained using centralized learning and distributed learning.
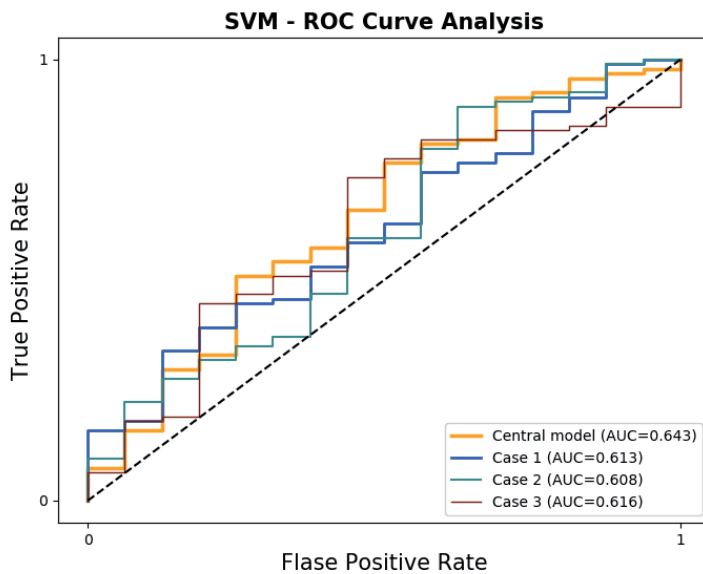
STAGE III NSCLC dataset

Figure S5.15: Receiver operator characteristic curves for overall survival model (SVM) trained using centralized learning and distributed learning.
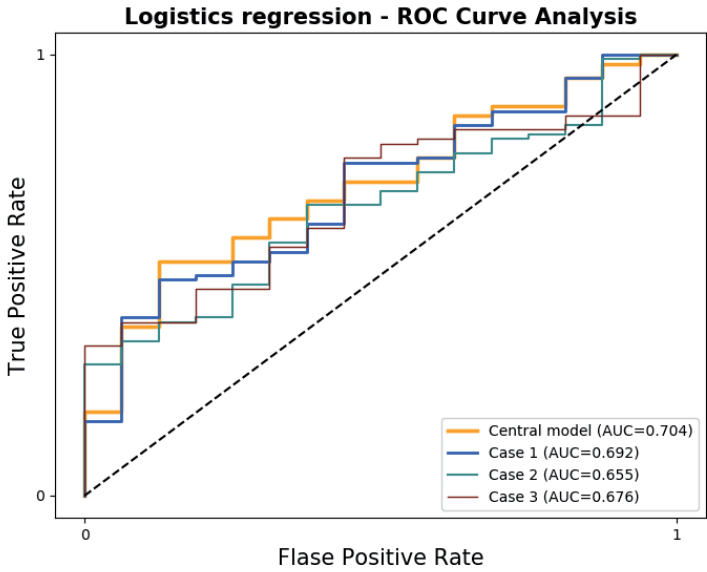
Figure S5.16: Receiver operator characteristic curves for overall survival model (Logistic regression) trained using centralized learning and distributed learning.
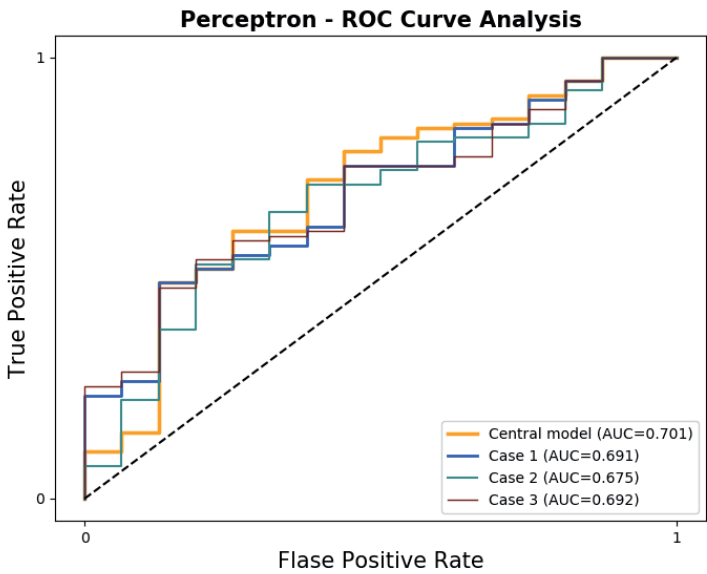

Figure S5.17: Receiver operator characteristic curves for overall survival model (Perceptron) trained using centralized learning and distributed learning.

Section C
Average difference in classifier AUC in percent

Table S5.1: Percentage difference and average difference in classifier AUC for each dataset (Breast cancer, Indian Liver, NSCLC-Radiomics dataset, and Stage III NSCLC)

Training scenario	Classifiers	Difference in AUC (%)			
		Breast cancer	Indian Liver	NSCLC-Radiomics dataset	Stage III NSCLC
Central model	SVM vs LR	1.1	0.00	12.5	9.37
	SVM vs Perceptron	0.00	2.63	12.5	9.37
	LR vs Perceptron	1.02	2.63	0.00	0.00
	Average difference	**0.67**	**1.75**	**8.33**	**6.24**
Case 1	SVM vs LR	2.02	1.29	10.93	13.11
	SVM vs Perceptron	1.01	1.29	6.25	13.11
	LR vs Perceptron	1.03	0.00	4.22	0.00
	Average difference	**1.35**	**0.86**	**7.13**	**8.74**
Case 2	SVM vs LR	1.02	1.35	10.76	6.66
	SVM vs Perceptron	1.02	0.00	3.07	11.66
	LR vs Perceptron	2.06	1.36	6.94	4.68
	Average difference	**1.36**	**0.90**	**6.92**	**7.66**
Case 3	SVM vs LR	1.02	1.33	12.90	9.83
	SVM vs Perceptron	1.02	4.00	6.45	13.11
	LR vs Perceptron	0.00	5.40	5.71	2.98
	Average difference	**0.68**	**3.57**	**8.35**	**8.64**

5

References

[1]   Y. Lecun, Gradient-Based Learning Applied to Document Recognition, PROCEEDINGS OF THE IEEE. 86 (1998) 47.

6

A blockchain based approach for Privacy Preserving distributed learning - Grade Group Prediction for Prostate Cancer Patients

Fadila Zerka | Visara Urovi | Akshayaa Vaidyanathan | Vikas Jaiman |
Ralph T.H. Leijenaar | Sean Walsh | Hanif Gabrani-Juma |
Henry C. Woodruff | Michel Dumontier | Philippe Lambin

Work in progress.

Abstract

Objectives: setting up centralized multicentric studies is a complex and time-consuming process. Distributed learning allows machine learning models to use data from multiple institutions without exchanging individual patient-level data. We demonstrate this technique in a cross institutional study to predict prostate cancer grade group of prostate cancer patients.

Materials and Methods: simulated a two-center distributed learning network to train and validate a prostate cancer grade group predictive model. The open-source dataset of Prostate MRI and Ultrasound with Pathology and Coordinates of Tracked Biopsy (n = 1151) was used to demonstrate the distributed learning framework.

Results: the model's performance was evaluated on a global test set of 192 subjects. The Global model showed an average AUC of 0.77 on the global test set for classification of five prostate cancer grade groups (Gleason score (GS) <= 6, GS = 3+4, GS = 4+3, GS = 8, and GS > 8).

Conclusion: distributed learning for classifying prostate cancer grade groups is feasible and presents a new methodology to analyze prostate cancer data without exposing any private information.

Introduction

Unlike other solid organ cancers, imaging is not used to diagnose patients with prostate cancer. Prostate cancer diagnosis is performed through biopsy for patients with a high level of prostate specific antigen (PSA) in their blood [1]. From the biopsy is estimated a Gleason Score which is a grading system used to rate the aggressiveness of prostate cancer and help choosing appropriate treatment options.

Artificial intelligence (AI) is a promising filed that has been successfully applied in medical imaging. AI was used for diagnosis and prognosis of different pathologies and diseases specially in radiology [2]. Similarly, AI approaches (i.e. machine learning and deep learning) have been implemented to prostate histopathology [3]. In this context, it was observed that machine learning applications for quantitative imaging (radiomics) as well as deep features can be used to predict Gleason score [4–6]. However, medical data is highly protected by law and ethics[7]. In consequence accessing and sharing such data became difficult and time consuming. Combining distributed learning and blockchain technology allows us to train the models without the data leaving their original locations [8]. The addition of a blockchain-based interaction provides the ability to trace back data provenance and usage of the created models. This combination prevents any data leakage and ensures the usage of the model for the purposes it was intended for initially.

We hypothesize that training of prostate cancer grade group prediction models in a distributed fashion across multiple sites is feasible and gives the similar performance than a centralized approach. Distributed learning provides an opportunity to train AI models with more data compared to data available at single centralized hospital. As this data originates from different centers; we expect our model to outperform models trained using data from a single center and to be more generalizable. Further the model can help predict prostate cancer grade group from prebiopsy magnetic resonance imaging (MRI).

In this paper, we propose a multicentric model for prostate cancer grade group prediction learned from multi-stage prostate cancer MRI of prostatectomy patients. The MRIs are coupled with both Biopsy and prostatectomy Gleason scores. Blockchain technology is used to validate and monetize the data used for training.

Materials and Methods

Data
We used the open source Prostate MRI and Ultrasound With Pathology and Coordinates of Tracked Biopsy (Prostate-MRI-US-Biopsy) dataset [9–11] to run this experiment. The

dataset consists of 1151 T2 weighted MRI observations of prostate cancer patients, each MRI is applied with prostate and tumor segmentation masks, and biopsy derived primary and secondary Gleason scores.

A total number of 14 observations were excluded from further analysis because they had missing segmentations, either prostate mask or tumor mask, or both, leaving 1110 eligible for prostate cancer grade group model training and global testing. Each 3D image was cropped around the prostate mask to make sure the prostate gland is available for model training. From each T2 weighted image we extracted three consecutive slices based on the diameter of the tumor. We firstly selected the slice containing the tumor with biggest diameter, we then selected the preceding and succeeding slices. Each slice is reshaped to a size of 64 by 64 pixels. The three consecutive slices are grouped as a three-channel input for the model. The data was split to training and testing sets (80% and 20% respectively). The training data was further split in to two sets (80% and 20% respectively) simulating two partners (Partner 1 and Partner 2). Prior training, each partner's set is split into 80% and 20% subsets for training and validation respectively. All data splits (Training, validation, and testing) were stratified based on the outcome label classes to guarantee equal percentage of positive and negative samples on each set. The primary and secondary Gleason scores were translated into labels containing five classes ranging from 0 to 4, as described in Table 6.1.

Table 6.1: Prostate cancer grade group, associated Gleason score, and labels

Grade group	Primary GS	Secondary GS	Combined GS	Labels
1	3	3	6	0
2	3	4	7	1
3	4	3	7	2
4	4,3,5	4,5,3	8	3
5	4,5,5	5,4,5	9 & 10	4

Real time data augmentation was performed on the final training sets (Partner 1 and Partner 2) to balance the different classes (0, 1, 2, 3, and 4). For each training set first determine the majority class, based on which the number the augmentation threshold is defined for the remaining classes. The augmentation threshold is the difference between the size of the majority class the each of the remaining classes' size, and it determines the number of augmented images that need to be produced to balance the data. The performed augmentation consists of clockwise rotation, anticlockwise rotation (with a random angle between 2 and 13 degrees respectively), random noise, and Gaussian blur.

Model architecture

The model consists of a ResNeXt-50 [12] pretrained on ImageNet dataset [13]. During the training process, the ADAM optimizer was applied with the learning rate (lr = 1e-5); model loss was calculated using categorical cross-entropy. The output of the model is classification probability of five categories.

Model training

We used the C-DistriM [8] software to set up a distributed learning network of two partners. Each partner had direct access patient-level data, the local learning and model parameter (such as the number of steps per epoch, number of epochs, and batch size) estimation is based on the data size. C-DistriM is a decentralized distributed learning framework based on Ethereum blockchain. C-DistriM's smart contract was deployed to Ethereum Rinkeby test network (https://www.rinkeby.io/) to enable the learning from local partner data.

Both partners data was placed in physical Surface Book 2 laptop (Intel(R) Core i7-8650U CPU, 16 GB RAM, equipped with NVIDIA GeForce GTX 1050 GPU), and running Windows 10 Pro. Ethereum blockchain was installed and synced using the official Go implementation of the Ethereum protocol Geth version 1.10.6-576681f2 (https://geth.ethereum. org/downloads/). An Ethereum account was created for each partner, the address of each partner and the smart contract can be found in Table 6.2, and the transaction history of each account can be accessed at the Rinkeby etherscan online platform (https://rinkeby.etherscan.io/).

Table 6.2: Corresponding Rinkeby test network addresses of the smart contract and partners (1 and 2)

Account holder	Address on Rinkeby test network
Smart contract	0xf3661ecd7e6c9c4bd8369462db1f7c1f87494363
Partner 1	0x1608076d8861c77feecc8a6e0f711776c63e8f4d
Partner 2	0x8541feebd9826d6ba4e13711278a1d43938d9795

6

The smart contract delegates the learning process by allowing only registered partners to proceed with the learning process. It provides the registered partner with an ERC271 Token to enable it to make the local model available to the other partner for update (more information on Blockchain technology, smart contracts and ERC271 Tokens are available in Supplementary Materials (Section A1). The following flowchart presents a detailed overview of the learning process:
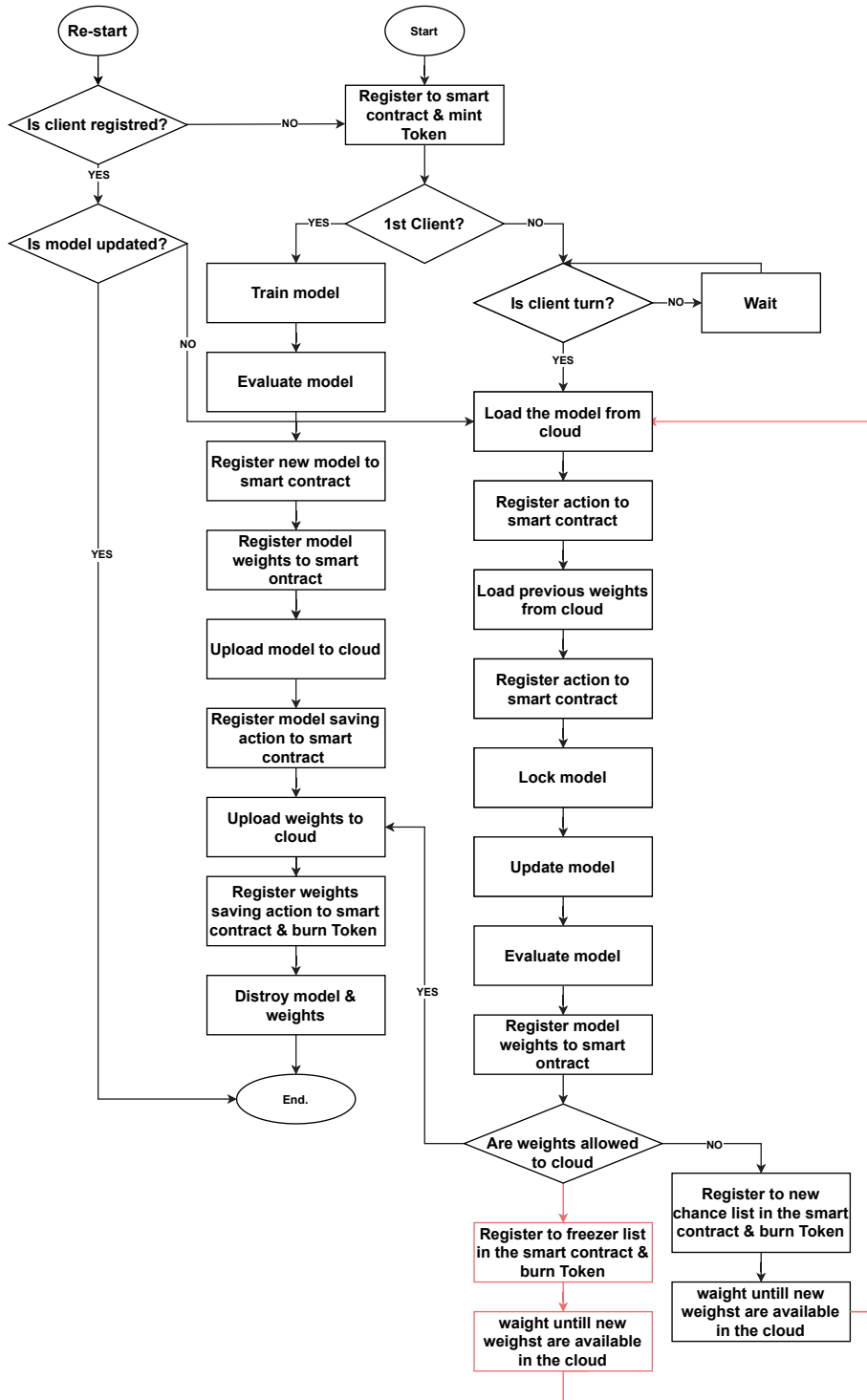
Figure 6.1: C-DistriM workflow.

Model performance in the local validation and global test set are expressed in term of the Area under the Curve (AUC) of the Receiver Operating Characteristic curve (ROC). The AUC is estimated for each class individually and the main AUC of the model is estimated by computing the average AUC of all possible pairwise combinations of the five classes, using the Python's scikit-learn package (v=0.24.2).

Results

A summary of cohort, tumor specificities in the dataset along with number of samples on each of the training, local validation, and global test sets is presented in Table 6.3.

Table 6.3: Dataset characteristics

			Partner 1 samples		Partner 2 samples		
Median age	Median PSA	Prostate volume	Training (after aug-mentation)	Validation	Training (after aug-mentation)	Validation	Global Testing samples
+65 (years)	4.9 ng/ml	48 cc	290	32	1135	108	192

The average AUC of the global model on the local validation set of the Partner 1 (first training round) and Partner 2 (second training round) is 0.63 and 0.73 respectively. The global model achieved an average AUC of 0.77 on the global test set. The ROC curves for each class obtained on the local validation and the global test sets for the global model are presented in Figure 6.2. The performance for prostate cancer grade group 1 (class 0) classification in the local validation set has an AUC of 0.54. Grade group 2 (class 1), grade group 3 (class 2), grade group 4 (class 3), grade group 5 (class 4) present an AUC of 0.42, 0.77, 0.77, and 0.77 respectively.

The performance for prostate cancer grade group 1 (class 0) classification in the global test set has an AUC of 0.63. Grade group 2 (class 1), grade group 3 (class 2), grade group 4 (class 3), grade group 5 (class 4) present an AUC of 0.55, 0.79, 0.79, and 0.79 respectively.

Discussion

This study demonstrates the feasibility of privacy preserving distributed learning for prostate cancer. We developed and tested a distributed deep learning model for prostate cancer grade group (aggressiveness status) classification based on biopsy derived Gleason scores. To demonstrate our approach we used the open-source Prostate-MRI-US-Biopsy dataset [9–11]. Our results showed an improvement of model performance
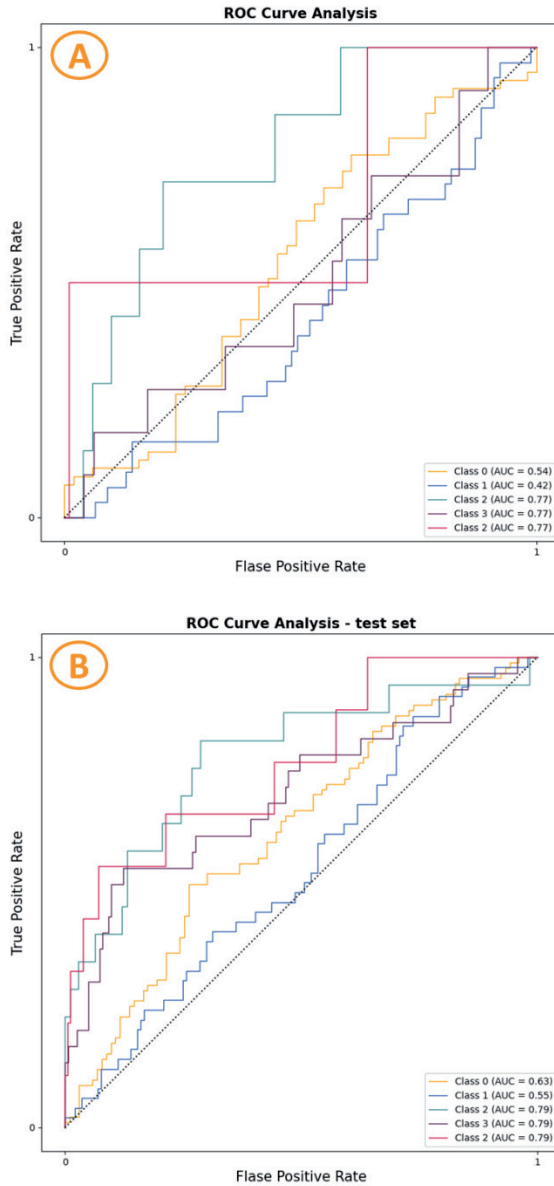
Figure 6.2: ROC curve (A) for local validation set (B) global test set.

in the global model compared to the model trained on the first partner's (Partner 1) data only. Distributed learning approaches were not previously investigated for prostate cancer grade group prediction. Herein, we present the potential of distributed learning to combine the knowledge retrieved from siloed MRI datasets.

At the present time the most widely used diagnosis method for prostate cancer is biopsy[14]. However, it may expose the patients to possible infections [15]. Imaging is conceived as a non-invasive approach for diagnosis and prognosis analysis of various cancers [16,17]. We leveraged the potential of MRI and deep learning to classify prostate cancer grading groups. Our model achieved an AUC of 0.73 on the local validation set and an AUC of 0.77 on the global test set for classification of prostate cancer grade groups based on Gleason score (GS <= 6, GS = 3+4, GS = 4+3, GS = 8, and GS > 8).

Several deep learning based studies were proposed to classify prostate cancer grade groups from MRI scans [18–21]. The results of these studies reported a classification AUC varying from 0.57 to 0.89. In spite of the fact that these studies achieved a good performance, the majority were trained on a single dataset limiting the model's capacity to generalize to other data collected using different acquisition parameters and protocols. In addition to that building a robust classification model requires an exposure to all the classification classes which is not the case in most of the datasets used in the proposed studies. Our distributed learning framework, however, provides the ability to update the classification models using new quality data (in terms of volume, veracity, variety, and velocity (4Vs) [22]) originating from different institutions. This exposure provides the model with an ability to generalize better. Moreover, class imbalance is known to affect the quality of the learning, to avoid overfitting on the majority class, we propose a data augmentation module that assesses on the fly how many new images should be generated for each class to balance the training on local dataset for each participating partner. Additionally, C-DistriM is based on Blockchain technology that ensures traceability and transparency of the learning process, thus the partners do not need to trust any third party for aggregating the learning.

Our study is limited to learning from biopsy labels, that are subject to errored estimations when compared to post-surgery labels [14]. Further analyses are required to evaluate training classification of prostate cancer grade groups using post-surgery labels and compare the model performance to the one obtained using biopsy labels. The data we used in this study is open source, our future work include expending the analysis presented in this work to learn from post-surgery data directly on the clinics while preserving the data privacy within the clinics' firewalls.

6

Conclusion

We have demonstrated the utility of privacy preserving distributed learning for classifying prostate cancer grade groups. We aim install our infrastructure in clinical environments and improve the performance and generalizability of our model.

Competing interests

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Fadila Zerka, Akshayaa Vaidyanathan, Fabio Bottari, Martin Gueuning, Hanif Gabrani-Juma, Mariaelena Occhipinti are salaried employees of Radiomics/Oncoradiomics SA. Dr Philippe Lambin reports, within and outside the submitted work, grants/sponsored research agreements from Varian medical, (Radiomics/Oncoradiomics), ptTheragnostic/DNAmito, Health Innovation Ventures. He received an advisor/presenter fee and/or reimbursement of travel costs/external grant writing fee and/or in kind manpower contribution from Radiomics/Oncoradiomics, BHV, Merck, Varian, Elekta, ptTheragnostic and Convert pharmaceuticals. Dr Lambin has shares in the company Radiomics/Oncoradiomics SA, Convert pharmaceuticals SA and The Medical Cloud Company SPRL and is co-inventor of two issued patents with royalties on radiomics (PCT/NL2014/050248, PCT/NL2014/050728) licensed to Radiomics/Oncoradiomics and one issue patent on mtDNA (PCT/EP2014/059089) licensed to ptTheragnostic/DNAmito, three non-patented invention (softwares) licensed to ptTheragnostic/DNAmito, Radiomics/Oncoradiomics and Health Innovation Ventures and three non-issues, non licensed patents on Deep Learning-Radiomics and LSRT (N2024482, N2024889, N2024889. Ralph T.H. Leijenaar has shares in the company Radiomics/Oncoradiomics and is co-inventor of an issued patent with royalties on radiomics (PCT/NL2014/050728) licensed to Radiomics/Oncoradiomics. Sean Walsh and Wim Vos have shares in the company Radiomics/Oncoradiomics. Michel Dumontier has shares in The Medical Cloud Company SPRL. Rest of the co-authors have no known competing financial interests or personal relationships to declare.

References

[1]   H.U. Ahmed, A. El-Shater Bosaily, L.C. Brown, R. Gabe, R. Kaplan, M.K. Parmar, Y. Collaco-Moraes, K. Ward, R.G. Hindley, A. Freeman, A.P. Kirkham, R. Oldroyd, C. Parker, M. Emberton, Diagnostic accuracy of multi-parametric MRI and TRUS biopsy in prostate cancer (PROMIS): a paired validating confirmatory study, The Lancet. 389 (2017) 815–822. https://doi.org/10.1016/S0140-6736(16)32401-1.

[2]   B. Sahiner, A. Pezeshk, L.M. Hadjiiski, X. Wang, K. Drukker, K.H. Cha, R.M. Summers, M.L. Giger, Deep learning in medical imaging and radiation therapy, Med. Phys. 46 (2019) e1–e36. https://doi.org/10.1002/mp.13264.

[3]   S.M. Ayyad, M. Shehata, A. Shalaby, M. Abou El-Ghar, M. Ghazal, M. El-Melegy, N.B. Abdel-Hamid, L.M. Labib, H.A. Ali, A. El-Baz, Role of AI and Histopathological Images in Detecting Prostate Cancer: A Survey, Sensors. 21 (2021) 2586. https://doi.org/10.3390/s21082586.

[4]   A. Chaddad, T. Niazi, S. Probst, F. Bladou, M. Anidjar, B. Bahoric, Predicting Gleason Score of Prostate Cancer Patients Using Radiomic Analysis, Front. Oncol. 8 (2018) 630. https://doi.org/10.3389/fonc.2018.00630.

[5]   A. Chaddad, M. Kucharczyk, T. Niazi, Multimodal Radiomic Features for the Predicting Gleason Score of Prostate Cancer, Cancers. 10 (2018) 249. https://doi.org/10.3390/cancers10080249.

[6]   K. Nagpal, D. Foote, Y. Liu, P.-H.C. Chen, E. Wulczyn, F. Tan, N. Olson, J.L. Smith, A. Mohtashamian, J.H. Wren, G.S. Corrado, R. MacDonald, L.H. Peng, M.B. Amin, A.J. Evans, A.R. Sangoi, C.H. Mermel, J.D. Hipp, M.C. Stumpe, Development and validation of a deep learning algorithm for improving Gleason scoring of prostate cancer, Npj Digit. Med. 2 (2019) 48. https://doi.org/10.1038/s41746-019-0112-2.

[7]   M.A. Foy, General Data Protection Regulation May 2018 (GDPR): how does it affect us?, 7 (2018) 3.

[8]   F. Zerka, V. Urovi, A. Vaidyanathan, S. Barakat, R.T.H. Leijenaar, S. Walsh, H. Gabrani-Juma, B. Miraglio, H.C. Woodruff, M. Dumontier, P. Lambin, Blockchain for Privacy Preserving and Trustworthy Distributed Machine Learning in Multicentric Medical Imaging (C-DistriM), IEEE Access. 8 (2020) 183939–183951. https://doi.org/10.1109/ACCESS.2020.3029445.

[9]   S. Natarajan, A. Priester, D. Margolis, J. Huang, L. Marks, Prostate MRI and Ultrasound With Pathology and Coordinates of Tracked Biopsy (Prostate-MRI-US-Biopsy), (2020). https://doi.org/10.7937/TCIA.2020.A61IOC1A.

[10]   G.A. Sonn, S. Natarajan, D.J.A. Margolis, M. MacAiran, P. Lieu, J. Huang, F.J. Dorey, L.S. Marks, Targeted Biopsy in the Detection of Prostate Cancer Using an Office Based Magnetic Resonance Ultrasound Fusion Device, Journal of Urology. 189 (2013) 86–92. https://doi.org/10.1016/j.juro.2012.08.095.

[11]   K. Clark, B. Vendt, K. Smith, J. Freymann, J. Kirby, P. Koppel, S. Moore, S. Phillips, D. Maffitt, M. Pringle, L. Tarbox, F. Prior, The Cancer Imaging Archive (TCIA): Maintaining and Operating a Public Information Repository, J Digit Imaging. 26 (2013) 1045–1057. https://doi.org/10.1007/s10278-013-9622-7.

[12]   S. Xie, R. Girshick, P. Dollár, Z. Tu, K. He, Aggregated Residual Transformations for Deep Neural Networks, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017: pp. 5987–5995. https://doi.org/10.1109/CVPR.2017.634.

[13]   J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, ImageNet: A Large-Scale Hierarchical Image Database, (n.d.) 8.

[14]   A. Camtosun, H. Gökçe, Comparison of prostate biopsy pathology and radical prostatectomy pathologies, Dicle Tıp Dergisi. (2019) 133–139. https://doi.org/10.5798/dicletip.534851.

[15]   L.V. Rodríguez, M.K. Terris, Risks and complications of transrectal ultrasound guided prostate needle biopsy: a prospective study and review of the literature, J Urol. 160 (1998) 2115–2120. https://doi.org/10.1097/00005392-199812010-00045.

[16]   H.J.W.L. Aerts, E.R. Velazquez, R.T.H. Leijenaar, C. Parmar, P. Grossmann, S. Carvalho, J. Bussink, R. Monshouwer, B. Haibe-Kains, D. Rietveld, F. Hoebers, M.M. Rietbergen, C.R. Leemans, A. Dekker, J. Quackenbush, R.J. Gillies, P. Lambin, Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach, Nat Commun. 5 (2014) 4006. https://doi.org/10.1038/ncomms5006.

[17]   P. Lambin, R.T.H. Leijenaar, T.M. Deist, J. Peerlings, E.E.C. de Jong, J. van Timmeren, S. Sanduleanu, R.T.H.M. Larue, A.J.G. Even, A. Jochems, Y. van Wijk, H. Woodruff, J. van Soest, T. Lustberg, E. Roelofs, W. van Elmpt, A. Dekker, F.M. Mottaghy, J.E. Wildberger, S. Walsh, Radiomics: the bridge between medical imaging and personalized medicine, Nat Rev Clin Oncol. 14 (2017) 749–762. https://doi.org/10.1038/nrclinonc.2017.141.

6

[18] Y. Yuan, W. Qin, M. Buyyounouski, B. Ibragimov, S. Hancock, B. Han, L. Xing, Prostate cancer classification with multiparametric MRI transfer learning model, Med Phys. 46 (2019) 756–765. https://doi.org/10.1002/mp.13367.

[19] R. Cao, A. Mohammadian Bajgiran, S. Afshari Mirak, S. Shakeri, X. Zhong, D. Enzmann, S. Raman, K. Sung, Joint Prostate Cancer Detection and Gleason Score Prediction in mp-MRI via FocalNet, IEEE Trans Med Imaging. 38 (2019) 2496–2506. https://doi.org/10.1109/TMI.2019.2901928.

[20] P. Khosravi, M. Lysandrou, M. Eljalby, M. Brendel, Q. Li, E. Kazemi, J. Barnes, P. Zisimopoulos, A. Sigaras, C. Ricketts, D. Meleshko, A. Yat, T.D. McClure, B.D. Robinson, A. Sboner, O. Elemento, B. Chughtai, I. Hajirasouliha, Biopsy-free prediction of prostate cancer aggressiveness using deep learning and radiology imaging, Radiology and Imaging, 2019. https://doi.org/10.1101/2019.12.16.19015057.

[21] S. Woo, S.Y. Kim, J.Y. Cho, S.H. Kim, Preoperative Evaluation of Prostate Cancer Aggressiveness: Using ADC and ADC Ratio in Determining Gleason Score, AJR Am J Roentgenol. 207 (2016) 114–120. https://doi.org/10.2214/AJR.15.15894.

[22] T. Lustberg, J. van Soest, A. Jochems, T. Deist, Y. van Wijk, S. Walsh, P. Lambin, A. Dekker, Big Data in radiation therapy: challenges and opportunities, The British Journal of Radiology. 90 (2017) 20160689. https://doi.org/10.1259/bjr.20160689.

Supplementary materials

Background

Blockchain
The idea of blockchain was initially proposed by Nakamoto et al. [1] as a digital cryptocurrency. It is a digital ledger that consists of transaction blocks and facilitates the transaction securely. Each block contains the hash of the previous block, as well as the timestamp. For public blockchain, the ledger is openly accessible for everyone but transactions once made and confirmed, are permanent and immutable. Nowadays, Blockchain technology has evolved to more general-purpose solutions such as Ethereum [2]. It uses a Proof of Work (PoW) consensus algorithm in which specific nodes, called *miners*, verify the validity of transactions. Once a transaction is confirmed, miners create a new block and append new valid blocks to the blockchain. To add the next block to the ledger, a *miner* must solve a complex mathematical problem and gets a reward in return for it. Moreover, due to the existence of smart contract functionality in Ethereum, it offers enormous functionalities that require agreements between two parties. Therefore, in our work, we use the Ethereum blockchain to interact between organizations.

Smart Contract
A smart contract is a digital protocol stored on an Ethereum blockchain that executes automatically when predefined conditions are met [3]. The main purpose of a smart contract is to translate contractual clauses between two parties into a code and enable automated workflows. It eliminates the need for a third party and increases efficiency. In this paper, we use smart contracts to define the conditions for training the models of different organizations.

Tokens
Ethereum tokens are a special type of cryptocurrency, usually defined as fungible, exchangeable assets [4]. The goal of using tokens is to provide a sophisticated system that allows automated access control to the data and serves as an identifier for specific operations to allow the supervisory authority to inspect these interactions. ERC-721 token [5] is a non-fungible asset i.e., the token cannot be swapped with another token and each token has a unique property. Non-fungible tokens (NFTs) represent data ownership over digital or physical assets and the ERC-721 standard provides functionality to track the NFTs. Therefore, in this work, we use the ERC-721 token to represent a unique key to the specific organization which allows us to differentiate over the data model used for learning.

6

References

[1]   S. Nakamoto, Bitcoin: A Peer-to-Peer Electronic Cash System, (n.d.) 9.

[2]   Ethereum Whitepaper, Ethereum.Org. (n.d.). https://ethereum.org (accessed July 26, 2021).

[3]   S. Szabo, The Idea of Smart Contracts | Satoshi Nakamoto Institute, (n.d.). https://nakamotoinstitute.
org/the-idea-of-smart-contracts/ (accessed July 26, 2021).

[4]   EIP-20: ERC-20 Token Standard, Ethereum Improvement Proposals. (n.d.). https://eips.ethereum.org/
EIPS/eip-20 (accessed October 27, 2020).

[5]   EIP-721: ERC-721 Non-Fungible Token Standard, Ethereum Improvement Proposals. (n.d.). https://
eips.ethereum.org/EIPS/eip-721 (accessed July 26, 2021).

6

7

Discussion

Distributed learning is a growing area of interest in healthcare and precision medicine, it enables access to data distributed over different medical institutions. Consequently, it facilitates big data in healthcare while preserving the privacy of the data and increasing the generalizability of the learned models. We described, in chapter 2, one of the major research fields that utilized distributed learning being the concept of radiomics and quantitative imaging in its conventional centralized settings. Furthermore, it described its challenges, pitfalls, and opportunities particularly data sharing and privacy preserving alternatives. We presented an overview of the challenges and opportunities attributed to distributed learning frameworks in chapter 3. We proposed, in chapter 4, a new decentralized distributed learning framework that addresses the challenges of conventional distributed learning frameworks presented in chapter 3. We then evaluated, in chapter 5, the impact of using decentralized distributed learning to learn from very small, siloed datasets. Furthermore, we evaluated our solution in an international setting, in chapter 6.

The following sections elaborate on the main challenges associated with distributed learning frameworks. Additionally, our contribution to overcome those challenges and future works will be discussed.

Multicentric Big data

Big data is a term employed to describe very large datasets, in healthcare domain it can be used for diagnostic and prognostic insight [1]. The significance of big data originates from the principle that large and heterogeneous datasets enable AI to learn and retain important characteristics better than small and homogenous datasets. In this context, issues related to data heterogeneity and variation are no longer relevant, as we illustrated in chapter 5, by simulating various distributed learning networks (n = 60000). However, a single healthcare organization cannot hold datasets that are big enough with respect to volume, veracity, variety, and velocity (4Vs) to be characterized as big data or to build reliable and generalizable AI [2]. Multicentric studies can leverage big data in healthcare and are particularly relevant for precision medicine initiative, early phase and pragmatic clinical trials. However, typical multicentric studies are slow and costly, primarily due to the legal and ethical considerations that hinders the process of data sharing and centralization in multicentric studies. Distributed learning offers a potential solution to these barriers by enabling fast and efficient multicentric studies while preserving the privacy of the data and ensuring similar performance as if the data was centralized [3,4].

7

Technical implementation

In the contrary to centralized learning processes described in chapter 2 for the use case of radiomics studies, technical implementation of distributed learning frameworks is a challenging process as it requires two main components: 1) <u>client</u>: which is the part of the software that is installed and is executed on a local machine owned or hosted by the hospital (or a research center) and on which the data is stored and accessible; 2) <u>server</u>: it consists of the main instance responsible to start, aggregate, and end the distributed learning process. Opensource frameworks have been released such as TensorFlow federated [5], Pysyft [6], Flower [7], and Vantage6 [8] to facilitate the implementation of federated learning frameworks for researchers. To utilize these frameworks, even if freely available, research centers invest considerable effort into integration and deployment using such frameworks. On the other hand (at the time of writing this thesis), there is a small number of companies (e.g., radiomics (Oncoradiomics SA) (chapter 6) and Owkin [9]) in addition to research groups (e.g., Dutch Personal Health Train [10,11], and Vantage 6 [12]), providing distributed learning solutions for medical research. These solutions are either production ready, built for long term use and technically validated or established for a fixed duration. However, since they are 1) only available as a commercial product or 2) dependent on a grant financing a research project for a predefined timeline, the impact on their use is limited, in the first case research community cannot fully benefit from the solution due to the lack of permanent financial support, and in the second case continuous development and integration are complex and hard to achieve. In both cases, the existing distributed learning frameworks need a server to allow the learning aggregation within a distributed network of medical institutions (example: hospitals, clinics, or research centers). This aggregation is an important step in distributed frameworks because it ensures the training continuity and safeguards the data used, model parameters/ weights, and the models learned. The aggregation task is ensured by a third-party server, to which the partners are connected and provide information concerning each step of the learning process. This leads to:

1. A centralization issue, as all the partners are required to trust the entity that holds the aggregation server.

2. A traceability issue, as the learning process is centralized, if the entity that manages the server is malicious it can falsity the records or not properly aggregate the learning. Detecting a server-side malignancy by the clients is very hard as they do not have any access to it.

3. The most common form of distributed learning is the one of "Federated learning" where the learning is parallel (all partners train local models simultaneously). This form of distributed learning even if it provides fast training and accurate results, it also represents a challenge in case a new partner joins

the network after the first execution of the project and if the related funding (case of academia) is extended and new data is pulled to the network. In this case the partners are subject to two main requirements technical and administrative, as it is necessary to rerun the analysis from scratch and contact all partners to submit new ethical approval forms and sign new agreements.

Challenges

Data access and manipulation

Data access from distributed learning frameworks is considerably different than data access in centralized settings. Data is typically stored and managed in different languages, following different guidelines of the electronic health records (EHR). This variation in the storing techniques makes the data access for distributed learning frameworks harder to achieve. In addition to these variations the researchers and developers of distributed learning frameworks cannot physically access or analyze the data during the preprocessing and learning processes [13]. In this matter, it is important to introduce standardization techniques for the data storage and usage, such as: 1) the digital imaging and communications in medicine (DICOM) standard [14], 2) the clinical data interchange standards consortium (CDISC) [15], 3) the health level seven (HL7) standards to guide data transfer and sharing [16], and FAIR (findable, accessible, interoperable, reusable) data principles [17], not only prior employment of distributed learning frameworks but as a routine practice, despite the type of the analysis that will be applied for the data. This standardization is key to effective manipulation of the data and saves time and expenses on the long term. Nonetheless, the integration of the aforementioned techniques is not straightforward and requires considerable efforts to setup and maintain.

Adoption & resistance to change

As discussed in chapter 3, the adoption distributed learning is subject to many impediments related technical, administrative, and informative constraints:

Accessing data in distributed learning settings requires a change or adaptation of a number of healthcare personnel to accomplish the corresponding tasks. In the present case, it is crucial to put considerable endeavor to facilitate and simplify the integration of the mentioned technologies in the healthcare domain. Avoiding positioning the personnel ahead of unknown and hard to interpret technologies or tools, may represent a step forward in embracing and adopting distributed learning frameworks by the healthcare partners.

Similarly with centralized frameworks, distributed learning frameworks are also subject to administrative manipulations such as grant writing and application for study ethical

7

and legal approvals. However, the lack of an in-depth knowledge of distributed learning methods makes the task of the ethical and legal departments more complex, and in parallel the adoption of such new technologies harder. It is important to document the integral distributed learning process by accentuating not only the data privacy concerns that are covered by distributed learning frameworks, but also the security measures taken to protect the data and its derivatives (such as the learned models and model weights), along with transparency and clearance measures to allow all the participants in such distributed frameworks monitor the learning process with no restrictions and trust concerns. This goal is very hard to achieve by the server-based frameworks, as the partners will always be required to entirely and blindly trust a single or multiple entities administrating the server in question. Thereby, establishing transparent decentralized distributed data frameworks is of major importance in easing the adoption of these technologies in healthcare and precision medicine.

Blockchain technology

Recently, blockchain has received great attention in the research community in healthcare and precision medicine. There are different types of blockchain technologies including public, private, and hybrid platforms, each one presents its own opportunities and challenges [18]. Herein, chapter 4 and 6, we focus on the public blockchains as no one has control over the information stored on it. Therefore, it is no longer necessary to trust an intermediary agent to mediate the learning process. This type of blockchains is based upon smart contracts, which are programming scripts that translate the legal agreements and terms agreed by all the healthcare partners involved in a network (distributed learning) into transactional actions. The smart contacts carry away the need for intermediaries and as such. Thus, it ensures process decentralization, enforces trust, immutability, transparency, traceability, and security.

Immutability

One of the major concerns associated with server-based distributed learning frameworks is process immutability. These systems are prone to falsification because they are managed by a single entity. Blockchain is immutable by design. It is built to safeguard all information stored in the ledger. It attains immutability through the use of cryptographic hashing, which derives from mapping data of arbitrary length to a fixed-size digest. Each transaction is registered after validation on a digital block. Each block includes a hash of the previous block, hence immutability. The immutability characteristic of blockchain provides the ability to enforce trust and grant integrity of the distributed learning frameworks amongst a network of healthcare partners.

Transparency

Blockchain runs on top of a consensus algorithm that enforces a transparent transaction mining process in the network. Moreover, since contracts deployed to the blockchain are public and in full view, they are permanently available for review and validation. Facilitating transparency of the distributed learning process enables the network partners to 1) monitor the learning process at any time during and after the training ends, 2) facilitate data governance in distributed learning projects, and 3) search any transaction amended to the chain during the learning.

Traceability

It is essential to provide process traceability to guarantee secure information sharing in distributed learning frameworks and encourage medical data providers to join such studies. Traceability enables real-time monitoring of data contribution to the learning process.

Network security

Blockchain, by means of its embedded immutability, is able to contribute to removing any threat related to partner misbehavior (such as providing falsified data). In addition, blockchain operates in combination with robust encryption protocols that ensure the security of the network.

Challenges and opportunities

The blockchain has shown its great potential within the research community. However, its adoption in clinical practice is yet to be determined.

- Public blockchains are subject to a scalability issue, for example Ethereum can mine approximately thirty transactions per second.

- Public blockchains have a single way of functioning. The entity that desires to integrate blockchain technology is responsible for restructuring its infrastructure to successfully coordinate both technologies.

- The transaction fee in public blockchains varies according to the network congestion status, whereas healthcare partners have fixed budgets for research projects. Therefore, it is complex to integrate a technology where it is hard to estimate the costs.

Blockchain is a rapidly evolving technology. New features are being implemented and released to improve and facilitate blockchain integration. For example, Ethereum is planning a new release (Ethereum 2, https://ethereum.org/en/eth2/) where the scalability is to be considerably improved (from thirty to thousands of transactions per second), consequently transaction fees will be reduced as well.

7

Technical considerations

As the main feature of public blockchains is to increase and enforce trust on the network, it is technically not suitable for some forms of distributed learning frameworks, such as federated learning, chapter 4. The core element of federated learning is weight aggregation. The weight aggregation procedure takes place in a server that is connected to each partners' site, where a local model is trained using local data. After completion of the learning, the local model weights are transferred to the server that averages all the local weights received from all or a part of the participating partners. Because of its reliance on an aggregation server, and because the current implementations of blockchain are not optimized to perform great computations such as weight averaging, there are substantial technical, computational, financial, and environmental challenges to overcome when attempting to integrate federated learning with blockchain. The most suitable form of distributed learning for integration with public blockchain is "sequential learning". In chapter 4 we have shown that integration of sequential learning and blockchain technology is feasible and that training predictive models sequentially delivers a similar result to the traditional centralized learning approach.

Distributed learning to increment model knowledge

In chapter 5 we discussed the value of distributed learning, performed in sequential settings, in incrementing the model knowledge in case of very small batch data. This type of learning is very beneficial in case of low prevalence diseases, where a single hospital may hold a single patient record or a small batch of patients only. As the application of AI improves with larger sets of data, it is unlikely a single medical center can train valuable machine learning models on low prevalence data. Sequential learning enables AI to learn new characteristics continuously and efficiently from medical data while maintaining privacy and model performance. The sequential functionality of this type of the learning does not only permit to continuously update the model when new data is available, but it also avoids repetitive administrative work, such as contacting all centers when a new center wants to join the network.

Federated learning frameworks, as well, demonstrated their potential in performing effective model learning in the case of small batch data [13]. The learning is held simultaneously (at the same time) on all the network partners. Consequently, the addition of a new partner requires contacting all the other partners and restarting the training from scratch. Under these circumstances, updating a federated learning model is conceived as a new study. Therefore, all legal and ethical applications must be carried out again.

In conclusion, sequential learning can significantly simplify the application of AI in case of law prevalence disease diagnosis. Nonetheless, the application of such frameworks, depends upon the efforts made to facilitate their integration on the clinical practice.

Future perspectives

The future vision of distributed learning is to reach a global learning scheme (Figure 7.1), where different types of data across the world can be made accessible to answer a variety of research questions. To achieve this vision, it is necessary to deliver improvements in three key areas of healthcare: technology, finance, and education.

Figure 7.1: Global clinical knowledge sharing.

In chapters 3 and 5 we integrated Ethereum, a public blockchain, to a distributed learning network to enforce traceability and address the trust issues related to the conventional distributed learning frameworks. Future works could be oriented to integrate the newest version of Ethereum, that addresses scalability and computational concerns in the current version. Moreover, as the transaction costs may be lower in the upcoming Ethereum version, more complex forms of sequential learning, such us updating the global model with local data in a cyclic manner at every epoch, have to be investigated.

In chapter 5, we validated the benefits of sequential distributed learning on small sets of radiomics and micro-array datasets. These types of datasets require feature selection prior to distributed learning in order to prevent overfitting [19,20]. Feature selection is yet to be investigated in the case of long-term sequential learning where model knowledge is incremented every time a new partner joins the distributed learning network. In this case, the raw data and features are extracted using different protocols and software, which leads to a variance in the robustness of certain features. Future work needs to investigate how to select common robust features if the data is not analyzed in one round.

7

National and international financial support is necessary to promote not only the implementation of distributed learning frameworks but the integration and the continuous development of these frameworks. Continuous development avoids repeating the work that has been accomplished previously by performing continuous, small, updates enabling the up-to-date code to be delivered to stakeholders as early as it is completed and tested. In addition to that, it is essential to bring research and industry together to benefit from 1) the industrial expertise in product setup, deployment, as well as manual and automated testing, 2) the research innovations in defining new AI methods to answer different research questions, and the invention of new pathways for distributed learning. The European Union started to support the combination of research and industry, a great example of financing projects is the Horizon 2020 project that provides support to research and industrial initiatives for distributed learning and other fields of research [21].

To improve the visibility and understanding of distributed learning concepts and increase the familiarity of healthcare providers with these concepts, more educational workshops need to be carried out. Hands-on workshops such as AI for imaging, both editions 2018 and 2019, are helpful to educate users how the process of carrying out a distributed learning process, and how distributed data analysis is implemented in a real life simulation [22]. High level courses, such as 3rd ESTRO Physics workshop [23], are also helpful to introduce distributed learning to the healthcare community by 1) increasing awareness of the distributed learning frameworks, 2) facilitating the integration of these frameworks to the healthcare information and communications technology (ICT) systems, 3) eliminating the need to trust a coordinating entity, and 4) prioritizing transparent traceability of the contribution and data, the adoption of distributed learning in healthcare will be feasible.

Conclusion

This thesis provides evidence of the validity of distributed multicentric studies as an alternative to centralized studies. We contributed by addressing the trust and traceability issues associated with conventional distributed learning frameworks. Continuous efforts are ongoing in improving distributed learning by means of performance and scalability. Nevertheless, additional efforts are required to answer all the technical questions related to distributed learning and facilitate the integration and adoption of such frameworks. We expect distributed learning in healthcare and precision medicine to be an active research area over the next decade and beyond.

References

[1]   S. Dash, S.K. Shakyawar, M. Sharma, S. Kaushik, Big data in healthcare: management, analysis and future prospects, J Big Data. 6 (2019) 54. https://doi.org/10.1186/s40537-019-0217-0.

[2]   T. Lustberg, J. van Soest, A. Jochems, T. Deist, Y. van Wijk, S. Walsh, P. Lambin, A. Dekker, Big Data in radiation therapy: challenges and opportunities, The British Journal of Radiology. 90 (2017) 20160689. https://doi.org/10.1259/bjr.20160689.

[3]   F. Zerka, V. Urovi, A. Vaidyanathan, S. Barakat, R.T.H. Leijenaar, S. Walsh, H. Gabrani-Juma, B. Miraglio, H.C. Woodruff, M. Dumontier, P. Lambin, Blockchain for Privacy Preserving and Trustworthy Distributed Machine Learning in Multicentric Medical Imaging (C-DistriM), IEEE Access. 8 (2020) 183939–183951. https://doi.org/10.1109/ACCESS.2020.3029445.

[4]   I. Feki, S. Ammar, Y. Kessentini, K. Muhammad, Federated learning for COVID-19 screening from Chest X-ray images, Applied Soft Computing. 106 (2021) 107330. https://doi.org/10.1016/j.asoc.2021.107330.

[5]   TensorFlow Federated, TensorFlow. (n.d.). https://www.tensorflow.org/federated (accessed June 21, 2021).

[6]   OpenMined/PySyft, OpenMined, 2021. https://github.com/OpenMined/PySyft (accessed June 21, 2021).

[7]   D.J. Beutel, T. Topal, A. Mathur, X. Qiu, T. Parcollet, P.P.B. de Gusmão, N.D. Lane, Flower: A Friendly Federated Learning Research Framework, ArXiv:2007.14390 [Cs, Stat]. (2021). http://arxiv.org/abs/2007.14390 (accessed October 31, 2021).

[8]   Home, (n.d.). https://distributedlearning.ai/ (accessed June 30, 2021).

[9]   Owkin Connect, our Federated Learning software, OWKIN. (n.d.). https://owkin.com/owkin-connect/ (accessed June 21, 2021).

[10]   The Personal Health Train Network | The Personal Health Train, (n.d.). https://pht.health-ri.nl/personal-health-train-network (accessed June 21, 2021).

[11]   T.M. Deist, F.J.W.M. Dankers, P. Ojha, M. Scott Marshall, T. Janssen, C. Faivre-Finn, C. Masciocchi, V. Valentini, J. Wang, J. Chen, Z. Zhang, E. Spezi, M. Button, J. Jan Nuyttens, R. Vernhout, J. van Soest, A. Jochems, R. Monshouwer, J. Bussink, G. Price, P. Lambin, A. Dekker, Distributed learning on 20 000+ lung cancer patients – The Personal Health Train, Radiotherapy and Oncology. 144 (2020) 189–200. https://doi.org/10.1016/j.radonc.2019.11.019.

[12]   A. Moncada-Torres, F. Martin, M. Sieswerda, VANTAGE6: an open source priVAcy preserviNg federaTed leArninG infrastructurE for Secure Insight eXchange, (n.d.) 8.

[13]   D. Ng, X. Lan, M.M.-S. Yao, W.P. Chan, M. Feng, Federated learning: a collaborative effort to achieve better medical imaging models for individual sites that have small labelled datasets, Quant Imaging Med Surg. 11 (2021) 852–857. https://doi.org/10.21037/qims-20-595.

[14]   DICOM, DICOM. (n.d.). https://www.dicomstandard.org (accessed August 5, 2021).

[15]   CDISC | Clear Data. Clear Impact., (n.d.). https://www.cdisc.org/ (accessed August 5, 2021).

[16]   Health Level Seven International - Homepage | HL7 International, (n.d.). https://www.hl7.org/ (accessed August 5, 2021).

[17]   M.D. Wilkinson, M. Dumontier, Ij.J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.-W. Boiten, L.B. da Silva Santos, P.E. Bourne, J. Bouwman, A.J. Brookes, T. Clark, M. Crosas, I. Dillo, O. Dumon, S. Edmunds, C.T. Evelo, R. Finkers, A. Gonzalez-Beltran, A.J.G. Gray, P. Groth, C. Goble, J.S. Grethe, J. Heringa, P.A.C. 't Hoen, R. Hooft, T. Kuhn, R. Kok, J. Kok, S.J. Lusher, M.E. Martone, A. Mons, A.L. Packer, B. Persson, P. Rocca-Serra, M. Roos, R. van Schaik, S.-A. Sansone, E. Schultes, T. Sengstag, T. Slater, G. Strawn, M.A. Swertz, M. Thompson, J. van der Lei, E. van Mulligen, J. Velterop, A. Waagmeester, P. Wittenburg, K. Wolstencroft, J. Zhao, B. Mons, The FAIR Guiding Principles for scientific data management and stewardship, Scientific Data. 3 (2016) 160018. https://doi.org/10.1038/sdata.2016.18.

[18]   F. Casino, T.K. Dasaklis, C. Patsakis, A systematic literature review of blockchain-based applications: Current status, classification and open issues, Telematics and Informatics. 36 (2019) 55–81. https://doi.org/10.1016/j.tele.2018.11.006.

[19]   S.M. Ayyad, A.I. Saleh, L.M. Labib, A new distributed feature selection technique for classifying gene expression data, Int. J. Biomath. 12 (2019) 1950039. https://doi.org/10.1142/S1793524519500396.

7

[20]   V. Bolón-Canedo, N. Sánchez-Maroño, A. Alonso-Betanzos, Distributed feature selection: An application to microarray data classification, Applied Soft Computing. 30 (2015) 136–150. https://doi.org/10.1016/j.asoc.2015.01.035.

[21]   kugleta, What is Horizon 2020?, Horizon 2020 - European Commission. (2013). https://ec.europa.eu/programmes/horizon2020/en/what-horizon-2020 (accessed June 23, 2021).

[22]   M. Akkers, HOME, Artificial Intelligence 4 Imaging. (n.d.). https://www.ai4imaging.org/ (accessed June 23, 2021).

[23]   V. Batista, H. Al-Hallaq, 3rd ESTRO Physics workshop: Science in development 25-26 October 2019, Budapest, Hungary, (2019) 2.

Appendices

A

Summary

Medical data is distributed across different health care providers and is difficult to share due to patient privacy protection regulations. In the last decade, the optimization of healthcare facilities and reaching a personalized level of clinical decision-making became highly dependent on the quantity and quality of the medical data. This data is used to learn useful and reliable artificial intelligence (AI), via machine or deep learning models. Ideally, data of all health care providers worldwide should be standardized and available to learn from. However, the current technical and financial resources, as well as legal and ethical considerations to protect patient data privacy, do not permit to centralize all data in a single database. In consequence, the implementation of AI applications in clinical practice is slowed down. New technical advancements such as distributed learning allow national and/ or multinational multicentric AI while preserving patient data privacy concerns.

In this thesis, we evaluated the current state of research in the field of radiomics. We then presented an up-to-date overview and limitations of distributed learning applications in health care. Also, we implemented a new infrastructure to address the limitations of the existing distributed learning solution. We studied the potential of decentralizing the learning process by enforcing trust, immutability, transparency, and traceability in distributed learning networks via blockchain smart contracts. Our results showed that the decentralization of distributed learning networks has the potential to train the models on patient data stored at different health care centers without need to exchange patient data and without need to trust a third party.

Furthermore, we evaluated the potential of distributed learning on small sets of data that can mimic the case of low prevalence diseases and phase I clinical trial, where it is not possible for a single health care provider to hold enough data to train reliable AI models. Our results suggest that distributed learning 1) overcomes data sharing limitations for low prevalence diseases and can be the new way to apply AI for such diseases, 2) is capable to train models as good as centrally trained models. Moreover, we deployed our infrastructure to a public blockchain (Ethereum) test network and tested the feasibility of decentralized distributed learning in real world settings. The decentralized nature of the study makes it easy for the public to evaluate the time, cost, and the learning quality, hence reaching transparent learning process.

Finally, this thesis showed that the decentralization of distributed learning processes is feasible and eliminates the trust issues associated with conventional distributed learning frameworks. Blockchain based decentralized distributed learning, thanks to its immutability, transparency, and traceability, may encourage health care providers to join distributed multicentric studies.

A

Nederlandse samenvatting

Medische gegevens is verspreid over verschillende zorgverleners en zijn moeilijk te delen vanwege de regels voor de bescherming van de privacy van patiënten. In het afgelopen decennium is de optimalisatie van zorginstellingen en het bereiken van een gepersonaliseerd niveau van klinische besluitvorming sterk afhankelijk geworden van de kwantiteit en kwaliteit van de medische gegevens. Deze gegevens worden gebruikt om bruikbare en betrouwbare kunstmatige intelligentie (AI) te trainen, d.m.v. machine- of deep learning-modellen. Idealiter zouden gegevens van alle zorgverleners wereldwijd gestandaardiseerd en beschikbaar moeten zijn om van te leren. De huidige technische en financiële middelen, evenals juridische en ethische overwegingen om de privacy van patiëntgegevens te beschermen, laten echter niet toe om alle gegevens in één enkele database te centraliseren. Hierdoor wordt de implementatie van AI-toepassingen in de klinische praktijk vertraagd. Nieuwe technische ontwikkelingen, zoals gedistribueerd leren, maken nationale en/of multinationale multicentrische AI mogelijk met behoud van bezorgdheid over de privacy van patiëntengegevens.

In dit proefschrift hebben we de huidige stand van het onderzoek op het gebied van radiomics geëvalueerd. Vervolgens hebben we een actueel overzicht en beperkingen van gedistribueerde leertoepassingen in de gezondheidszorg gepresenteerd. We hebben ook een nieuwe infrastructuur geïmplementeerd om de beperkingen van de bestaande gedistribueerde leeroplossing aan te pakken. We hebben het potentieel bestudeerd van decentralisatie van het leerproces door vertrouwen, onveranderlijk-heid, transparantie en traceerbaarheid in gedistribueerde leernetwerken af te dwingen via slimme blockchain-contracten. Onze resultaten toonden aan dat de decentralisatie van gedistribueerde leernetwerken het potentieel heeft om de modellen te trainen op patiëntgegevens die zijn opgeslagen in verschillende zorgcentra zonder dat patiënt-gegevens uitgewisseld hoeven te worden én zonder dat de gegevens aan een derde partij moeten te worden vertrouwd.

Verder evalueerden we het potentieel van gedistribueerd leren op kleine gegevenssets die ziekten met een lage prevalentie en fase I klinische onderzoeken kunnen nabootsen, waar het voor een enkele zorgverlener niet mogelijk is om over voldoende gegevens te beschikken om betrouwbare AI-modellen te trainen. Onze resultaten suggereren dat gedistribueerd leren 1) de beperkingen van het delen van gegevens voor ziekten met een lage prevalentie overwint en de nieuwe manier kan zijn om AI voor dergelijke ziekten toe te passen, 2) in staat is om modellen even goed als centraal getrainde modellen te trainen. Bovendien hebben we onze infrastructuur geïmplementeerd in een openbaar blockchain-testnetwerk (Ethereum) en hebben we de haalbaarheid van gedecentraliseerd gedistribueerd leren in de echte wereld getest. Het gedecentraliseerde karakter van het onderzoek maakt het voor het publiek gemakkelijk om de tijd, de kosten en de leerkwaliteit te evalueren, waardoor een transparant leerproces wordt bereikt.

Ten slotte toonde dit proefschrift aan dat decentralisatie van gedistribueerde leerprocessen haalbaar is en de vertrouwensproblemen elimineert die samenhangen met conventionele gedistribueerde leerkaders. Op blockchain gebaseerd gedecentraliseerd gedistribueerd leren kan, dankzij de onveranderlijkheid, transparantie en traceerbaarheid ervan, zorgverleners aanmoedigen om deel te nemen aan gedistribueerde multicentrische onderzoeken.

A

Impact paragraph

Artificial intelligence demands a large amount of high-quality data to build reliable and generalizable models. Conventionally data analysis models are learned using centralized datasets. In a centralized setting, data is collected and shared to one database. Data analysis models such as machine/deep learning learn directly from the data. Data sharing and centralization is however strictly regulated by legal and ethical considerations making the process of publicly sharing the data time consuming and costly.

Distributed and federated learning infrastructures allow to apply machine/deep learning algorithms, in addition to other statistical data analysis in multicentric settings without patient data and information ever leaving the individual clinical institutes. These infrastructures are designed to preserve patient data privacy. These infrastructures however suffer from luck of traceability and trust. Despite that, to apply these infrastructures in personalized medicine, and for them to last and be part of the medical device development process, the following points need to be addressed: 1) intensify traceability of patient data, 2) boost the trust amongst distributed learning partners, 3) provide evidence regarding the added value of distributed learning on providing optimal treatment options for individual patients.

Peer-to-peer distributed frameworks such as blockchain (Ethereum in our study), provide a secure decentralized ledger that allows individuals to securely transfer and exchange value. Ethereum is an open-source public blockchain, that enables smart contact development. The computer programs of the smart contracts enforce the execution of the terms of legal agreements (contracts).

In our study, we integrated blockchain to distributed learning and proposed a fully decentralized infrastructure. The integration of Blockchain to distributed learning infrastructures is of high importance as it 1) eliminates the need to trust a third parity that aggregates the learning process (fully decentralizes the learning process), 2) provides means to trace partner involvement in the learning process (insured by the immutability of the transaction history), and 3) sets a transparent environment for the commercialization and intellectual property agreements related to the medical devises learned in a distributed fashion.

Scientific impacts
1. All our studies are published in medical and technical journals (such as: Medicinal Research Reviews, JCO Clinical Cancer Informatics, IEEE Access, and Computers in Biology and Medicine).

2. All our studies are open access.

3. The main contribution of this thesis is the application of blockchain technology to distributed learning infrastructures (decentralized learning). The studies carried out throughout this thesis are a baseline and starting point to investigate multiple diagnosis and prognosis clinical and/or imaging analysis. Additionally, decentralized learning, by exposing the model to more data, can help in improving the predictive value of existing and new models such as radiomics signatures as well as automated segmentation models. This in turn will help to reduce the time of the decision-making process.

4. **Chapter 2** presents an overview of radiomics (as one of the research fields that utilized distributed learning), assessed its challenges, pitfalls, and future directions.

5. **Chapter 3** appraises the existing implementations of conventional distributed learning frameworks and discusses possible future pathways.

6. **Chapter 4** showed that utilization blockchain technology for distributed learning is feasible and is essential to decentralize the learning process, enforce trust, immutability, transparency, traceability, and security.

7. **Chapter 5** in dictated that distributed learning is feasible and promising and in case of the diseases with a low prevalence.

8. **Chapter 6** evaluated the infrastructure proposed in chapter 4 in a distributed learning network.

Social impacts

1. Distributed learning, especially decentralized learning, has the potential to expose predictive models to more data, hence, help in improving the predictive value of existing and new radiomics signatures as well as the automated segmentation models. This in turn will help to reduce the time of decision-making process and reduce medical practitioners' workload.

2. Continuous improvement of diagnosis and prognosis tools can improve patient care systems and reduce mortality rates.

3. Fast and accurate predictive models may save time and reduce costs for patients.

A

Curriculum vitae

Fadila Zerka was born on the 09th of June 1991 in Bejaia, Algeria. After high school at Lycée mixte D'adekar in Bejaia in 2010, she studied Applied Mathematics at the University of Abderrahmane Mira in Bejaia, Algeria. She performed two internships: first at Bejaia Mediterranean Terminal which focused optimizing the manipulation of containers, second at Division de Distribution de l'Electricité et du Gaz de l'Est which focused on developing a model for detection of the vulnerable nodes in a gas network. After obtaining her bachelor's degree in 2013, she was granted a one-year exchange to study computer science and information technology (CSIT) at the Technical University of Lodz in Poland. During the exchange year, she investigated the potential of differential differentiation for edge detection and compared it with other existing techniques. After that, she enrolled in the master's program of CSIT at the Technical University of Lodz in Poland. In parallel with her master studies, she worked as an engineer at Fujitsu IT Solutions. After her graduation, she carried out her position at Fujitsu IT Solutions, where she gained an interest in research. In 2018, she started a PhD project in a joint program between Radiomics in Belgium and Maastricht University in the Netherlands. Her PhD focused on Distributed learning for optimal Radiomics knowledge. During her PhD, she visited the laboratoire de traitement de l'information médicale (LATIM) at the University of Western Brittany, brest, France for a period of three months, to learn about radiomics feature harmonization. She also remotely joined the D-LAB at Maastricht University for three months to test her distributed learning infrastructure.

A

List of manuscripts

Published

F. Zerka, V. Urovi, F. Bottari, R.T.H. Leijenaar, S. Walsh, H. Gabrani-Juma, M. Gueuning, A. Vaidyanathan, W. Vos, M. Occhipinti, H.C. Woodruff, M. Dumontier, P. Lambin, Privacy preserving distributed learning classifiers – Sequential learning with small sets of data, Computers in Biology and Medicine. 136 (2021) 104716. https://doi.org/10.1016/j.compbiomed.2021.104716.

F. Zerka, V. Urovi, A. Vaidyanathan, S. Barakat, R.T.H. Leijenaar, S. Walsh, H. Gabrani-Juma, B. Miraglio, H.C. Woodruff, M. Dumontier, P. Lambin, Blockchain for Privacy Preserving and Trustworthy Distributed Machine Learning in Multicentric Medical Imaging (C-DistriM), IEEE Access. 8 (2020) 183939–183951. https://doi.org/10.1109/ACCESS.2020.3029445.

F. Zerka, S. Barakat, S. Walsh, M. Bogowicz, R.T.H. Leijenaar, A. Jochems, B. Miraglio, D. Townend, P. Lambin, Systematic Review of Privacy-Preserving Distributed Machine Learning From Federated Databases in Health Care, JCO Clinical Cancer Informatics. (2020) 184–200. https://doi.org/10.1200/CCI.19.00047.

J. Guiot, A. Vaidyanathan, L. Deprez, F. Zerka, D. Danthine, A. Frix, P. Lambin, F. Bottari, N. Tsoutzidis, B. Miraglio, S. Walsh, W. Vos, R. Hustinx, M. Ferreira, P. Lovinfosse, R.T.H. Leijenaar, A review in radiomics: Making personalized medicine a reality via routine imaging, Med Res Rev. (2021) med.21846. https://doi.org/10.1002/med.21846.

J. Guiot, A. Vaidyanathan, L. Deprez, F. Zerka, D. Danthine, A.-N. Frix, M. Thys, M. Henket, G. Canivet, S. Mathieu, E. Eftaxia, P. Lambin, N. Tsoutzidis, B. Miraglio, S. Walsh, M. Moutschen, R. Louis, P. Meunier, W. Vos, R. Leijenaar, P. Lovinfosse, Development and validation of an automated radiomic CT signature for detecting COVID-19, Infectious Diseases (except HIV/AIDS), 2020. https://doi.org/10.1101/2020.04.28.20082966.

A. Vaidyanathan, M.F.J.A. van der Lubbe, R.T.H. Leijenaar, M. van Hoof, F. Zerka, B. Miraglio, S. Primakov, A.A. Postma, T.D. Bruintjes, M.A.L. Bilderbeek, H. Sebastiaan, P.F.M. Dammeijer, V. van Rompaey, H.C. Woodruff, W. Vos, S. Walsh, R. van de Berg, P. Lambin, Deep learning for the fully automated segmentation of the inner ear on MRI, Sci Rep. 11 (2021) 2885. https://doi.org/10.1038/s41598-021-82289-y.

W. Rogers, S. Thulasi Seetha, T.A.G. Refaee, R.I.Y. Lieverse, R.W.Y. Granzier, A. Ibrahim, S.A. Keek, S. Sanduleanu, S.P. Primakov, M.P.L. Beuque, D. Marcus, A.M.A. van der Wiel, F. Zerka, C.J.G. Oberije, J.E. van Timmeren, H.C. Woodruff, P. Lambin, Radiomics: from qualitative to quantitative imaging, BJR. 93 (2020) 20190948. https://doi.org/10.1259/bjr.20190948.

In preparation/submitted

F. Zerka, V. Urovi, A. Vaidyanathan, R.T.H. Leijenaar, S. Walsh, H. Gabrani-Juma, H.C. Woodruff, M. Dumontier, P. Lambin, A blockchain based approach for Privacy Preserving distributed learning - Grade Group Prediction for Prostate Cancer Patients. (In preparation)

A. Vaidyanathan, J. Guiot, F. Zerka, F. Belmans, I. Van Peufflik, L. Deprez, D. Danthine, G. Canivet, P. Lambin, S. Walsh, M. Occchipinti, P. Meunier, W. Vos, P. Lovinfosse, R.T.H. Leijenaar, An externally validated fully automated deep learning algorithm to classify COVID-19 and other pneumonias on chest CT. (Submitted)

A