IEEE Access
Multidisciplinary : Rapid Review : Open Access Journal

Received September 11, 2020, accepted September 24, 2020, date of publication October 7, 2020, date of current version October 20, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3029445

Blockchain for Privacy Preserving and Trustworthy Distributed Machine Learning in Multicentric Medical Imaging (C-DistriM)

FADILA ZERKA[1,2], VISARA UROVI[3], AKSHAYAA VAIDYANATHAN[1,2], SAMIR BARAKAT[2], RALPH T. H. LEIJENAAR[2], SEAN WALSH[1,2], HANIF GABRANI-JUMA[2], BENJAMIN MIRAGLIO[2], HENRY C. WOODRUFF[1,4], MICHEL DUMONTIER[3], AND PHILIPPE LAMBIN[1,4]

[1]The D-Lab, Department of Precision Medicine, GROW – School for Oncology, Maastricht University, 6229 GT Maastricht, The Netherlands
[2]Oncoradiomics SA, 4000 Liège, Belgium
[3]Institute of Data Science (IDS), Maastricht University, 6229 GT Maastricht, The Netherlands
[4]Maastricht University Medical Centre, Department of Radiology and Nuclear Medicine, 6202 AZ Maastricht, The Netherlands

Corresponding author: Fadila Zerka (fadila.zerka@oncoradiomics.com)

This work was supported by ERC advanced Grant (ERC-ADG-2015, n° 694812 - Hypoximmuno). This research is also supported by the Dutch Technology Foundation STW (Grant n° P14-19 Radiomics STRaTegy), which is the applied science division of NWO, Aspasia NWO (Grant n°91716421) and the Technology Program of the Ministry of Economic Affairs. Authors also acknowledge financial support from SME Phase 2 (RAIL - n° 673780), EUROSTARS (DART - n° E10116, DECIDE - n° E11541), the European Program PREDICT - ITN - n° 766276), TRANSCAN Joint Transnational Call 2016 (JTC2016 "CLEARLY" - n° UM 2017-8295) and Interreg V-A Euregio Meuse-Rhine ("Euradiomics" - n° EMR4).

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-val > 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.

INDEX TERMS Blockchain, data privacy, decentralized learning, distributed learning.

I. 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

The associate editor coordinating the review of this manuscript and approving it for publication was Dongxiao Yu.

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.

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.

II. BACKGROUND AND SIGNIFICANCE

A. 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].

B. 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.

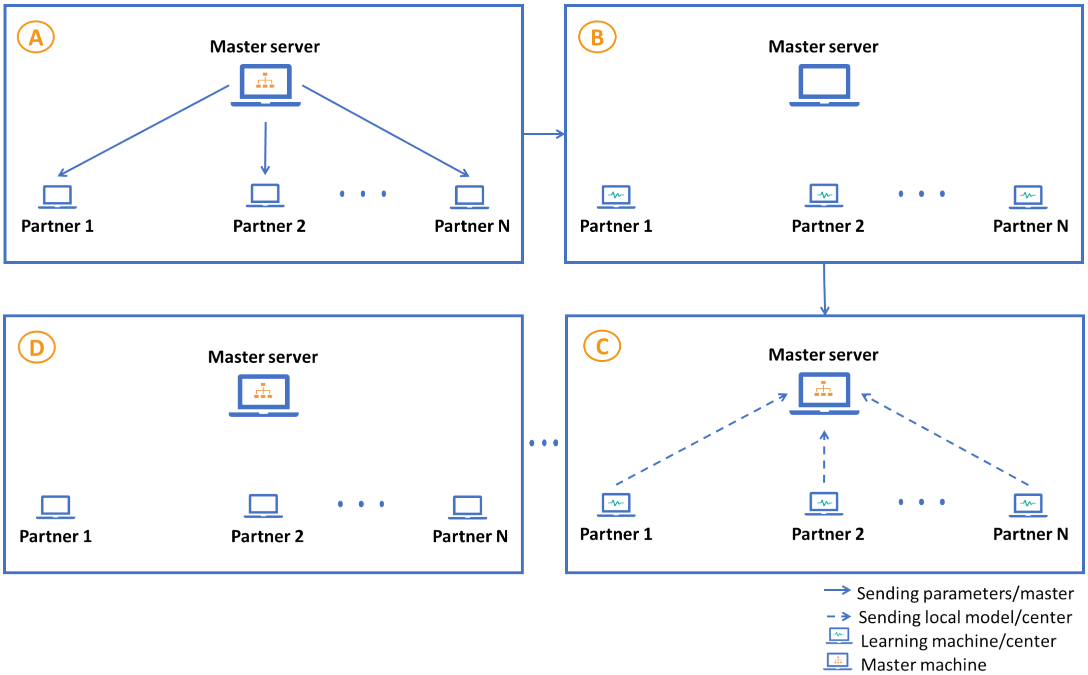

FIGURE 1. Conventional distributed learning (federated learning) process, (A) Master server initialize the learning by sending initial models to the partners (B) partners train the received model with local data, (C) partners send updated models to the master server, (D) master server aggregates the received models and verify convergence criteria.

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 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.

C. 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.

III. MATERIALS AND METHODS

A. 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 1.

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

B. CHAINED DISTRIBUTED MACHINE LEARNING (C-DISTRIM)

1) ARCHITECTURE

The objective of C-DistriM is to train distributed models with equivalent performance as centralized models, preserve

TABLE 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 %

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 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.
- 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.

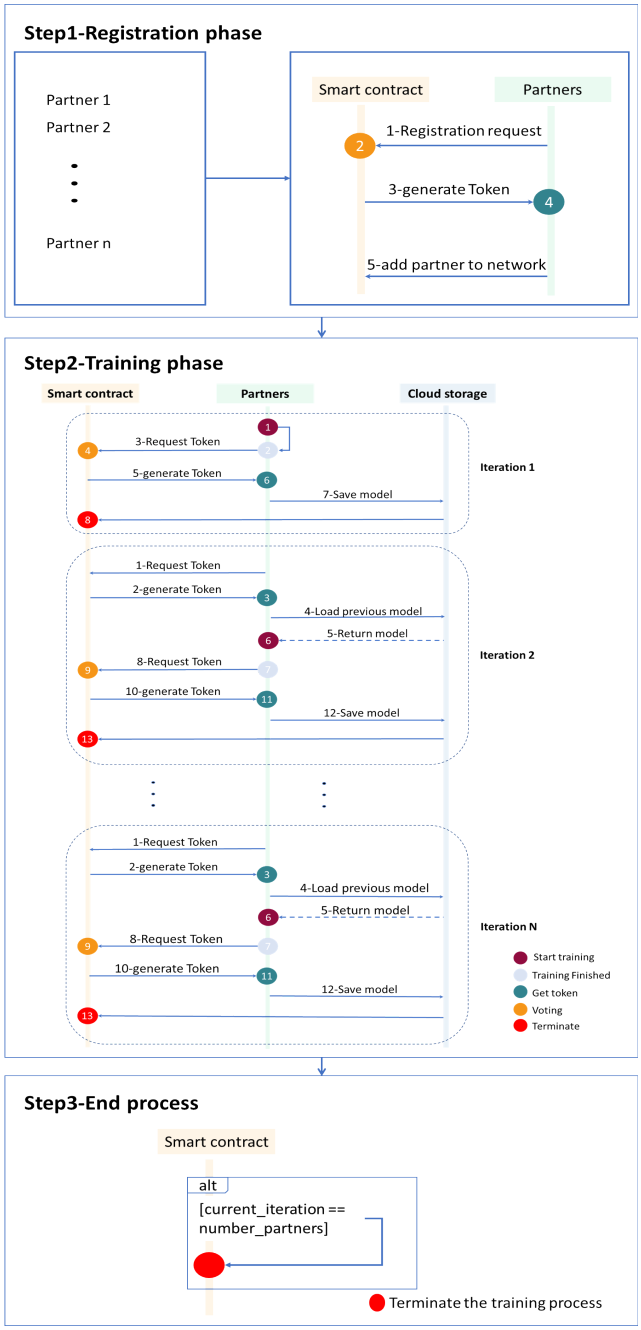

FIGURE 2. Overview of C-DistriM: (step 1) partners register to the network through the smart contract; (step 2) training starts by iterating through the partner list: (1) start training the first local model; (2) when training ends; (3) request a token to save the model to cloud; (4) vote to decide if 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.

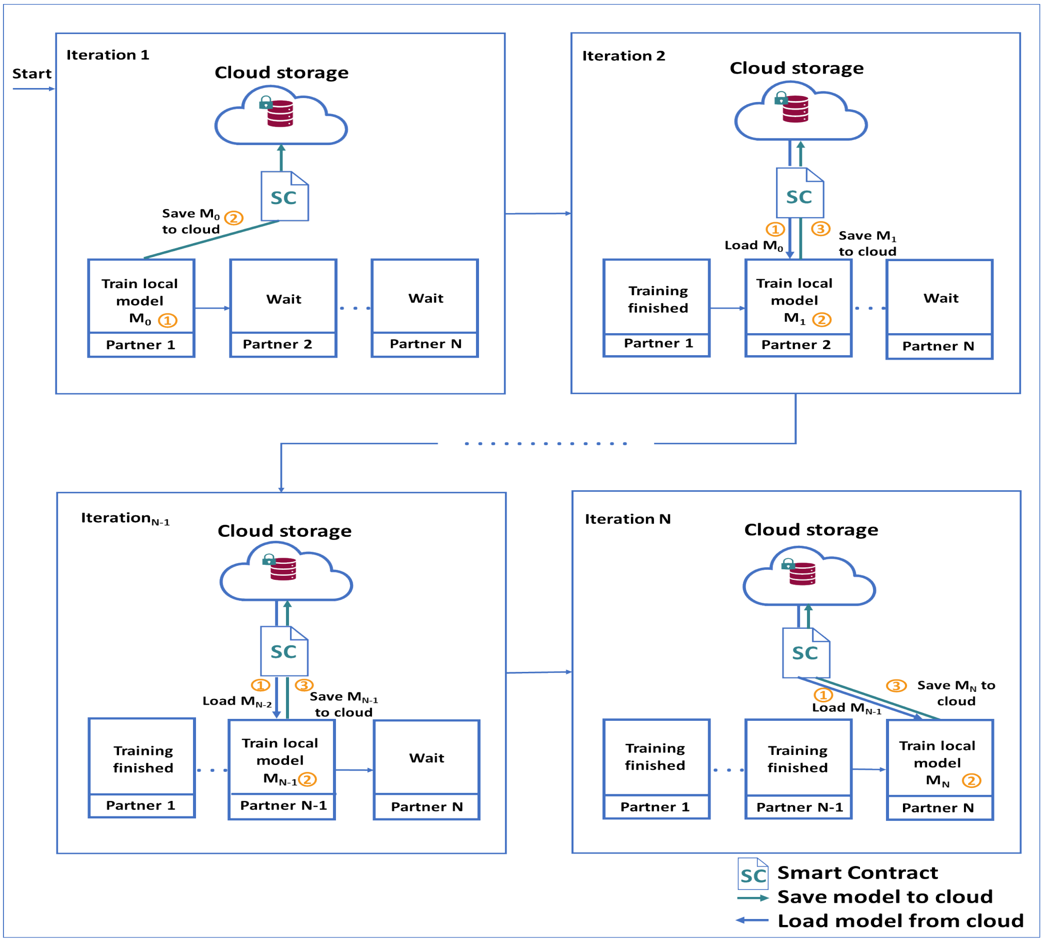

FIGURE 3. Distributed learning flow diagram, each iteration corresponds to one partner update of the model with local data.

2) 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 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 platform for testing purposes. This means that our prototype is currently tested with a local testnet, 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.

C. TRAINING

1) 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 appendix A.

2) 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 3 × 3 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 × 160 × 16) pixels for model training and validation process.

3) 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.

IV. 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 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. 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.

From the calibration plots presented in FIGURE 5, we can observe a variation in the calibration of the models.

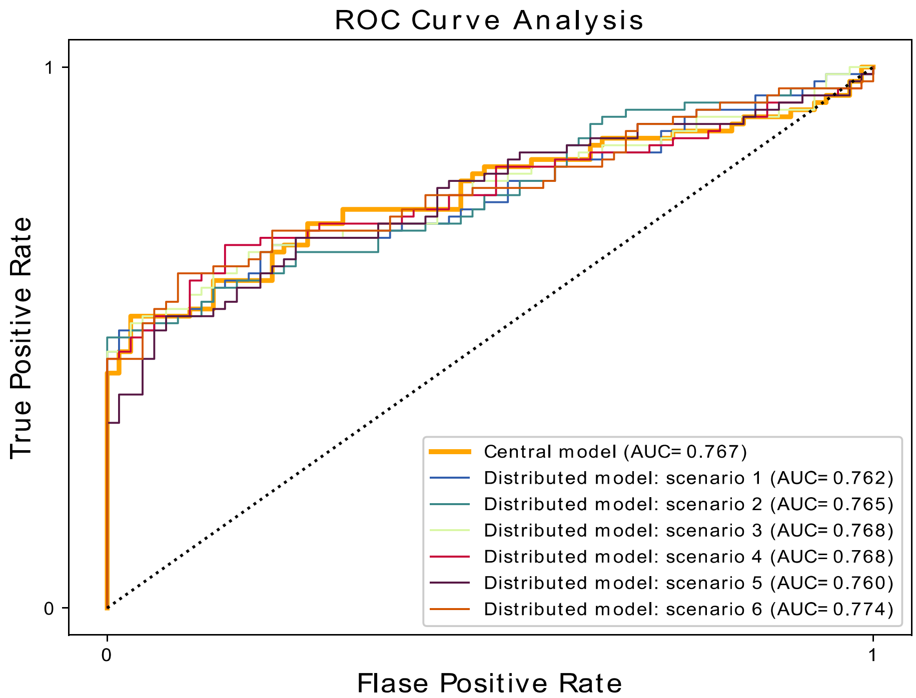

FIGURE 4. Receiver operator characteristic curves for two-year survival model trained using centralized learning and distributed learning.

TABLE 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)

The IRIS conclusions were the same as the NSCLC-Radiomics dataset use-case. Detailed results are presented in appendix A.

Appendix 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.

V. 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

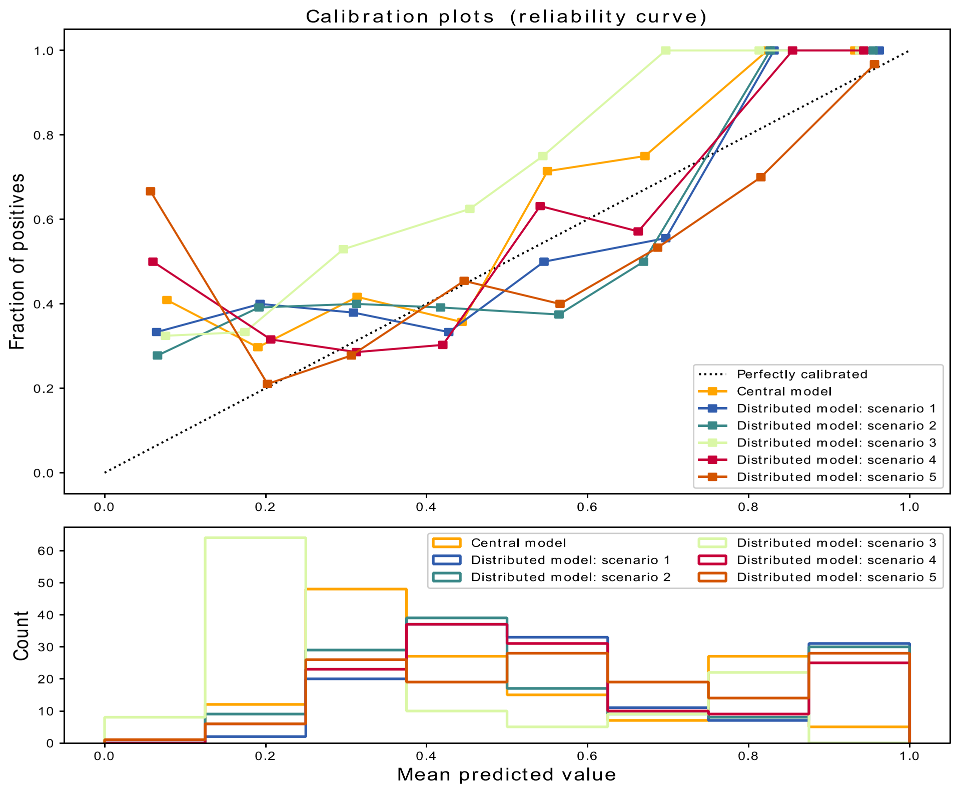

FIGURE 5. **Calibration curves for centralized and distributed learning models.**

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[1]. 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 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

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

TABLE 3. Summary of different blockchain infrastructures used for privacy preserving distributed learning.

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)

"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 3 illustrates the key differences between the listed blockchain infrastructures and the proposed C-DistriM.

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.

VI. 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.

APPENDIX A

A. MATERIALS AND METHODS

1) 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.

2) 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.

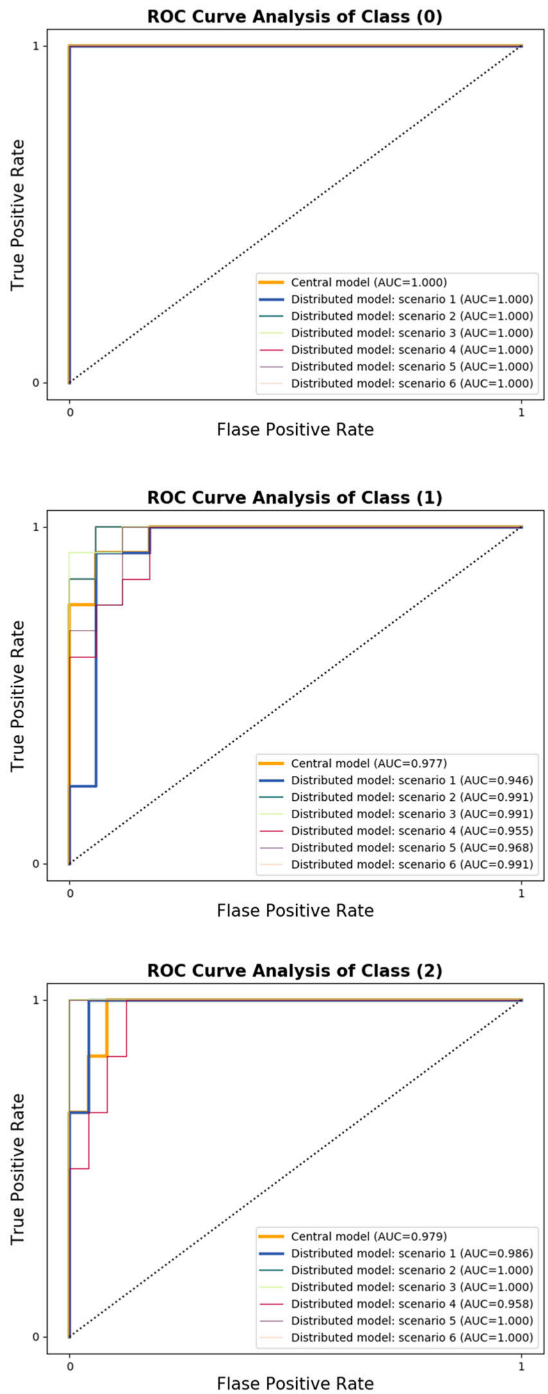

FIGURE 6. **Receiver operator characteristic curves for IRIS species prediction model trained using centralized learning and distributed learning.**

B. RESULTS

The ROC curves for each class are illustrated in FIGURE 6, and the comparative performance of each approach (in 95 % confidence interval) is illustrated in Table 4.

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.

TABLE 4. **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)

APPENDIX B

See Fig. 7.

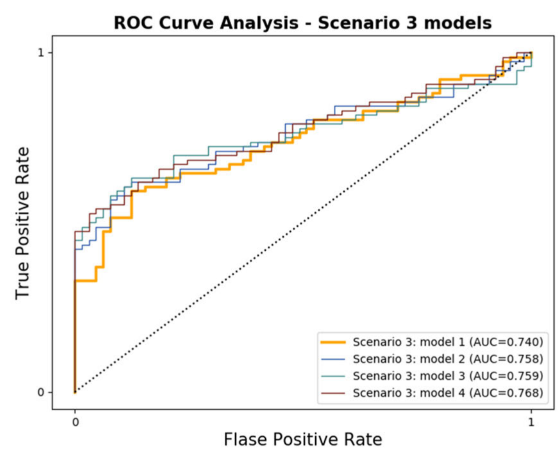

FIGURE 7. **Receiver operator characteristic curves for two-year survival models trained using "Scenario 3" data distribution, (A) model 1: data from center 1 only (n = 141); (B) model 2: data from center 1 and center 2 (n = 141 + n = 140); (C) data from center 1, center 2, center 3 (n = 141 + n = 140 + n = 141); (D) data from center 1, center 2, center 3, center 4 (n = 141 + n = 140 + n = 141 + n = 141).**

ACKNOWLEDGMENT

The authors thank Dr. W. Vos, F. Bottari, and S. Moorman for editing the manuscript.

REFERENCES

[1] S. Kulkarni, N. Seneviratne, M. S. Baig, and A. H. A. Khan, "Artificial intelligence in medicine: Where are we now?" *Acad. Radiol.*, vol. 27, no. 1, pp. 62–70, Jan. 2020, doi: 10.1016/j.acra.2019.10.001.

[2] Y. Zhou, J. Xu, Q. Liu, C. Li, Z. Liu, M. Wang, H. Zheng, and S. Wang, "A radiomics approach with CNN for shear-wave elastography breast tumor classification," *IEEE Trans. Biomed. Eng.*, vol. 65, no. 9, pp. 1935–1942, Sep. 2018, doi: 10.1109/TBME.2018.2844188.

[3] X. Jiang, J. Li, Y. Kan, T. Yu, S. Chang, X. Sha, H. Zheng, Y. Luo, and S. Wang, "MRI based radiomics approach with deep learning for prediction of vessel invasion in early-stage cervical cancer," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, early access, Jan. 3, 2020, doi: 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, and S. Walsh, "Radiomics: The bridge between medical imaging and personalized medicine," *Nature Rev. Clin. Oncol.*, vol. 14, no. 12, pp. 749–762, Dec. 2017, doi: 10.1038/nrclinonc.2017.141.

[5] M. Owais, M. Arsalan, J. Choi, and K. R. Park, "Effective diagnosis and treatment through content-based medical image retrieval (CBMIR) by using artificial intelligence," *J. Clin. Med.*, vol. 8, no. 4, p. 462, Apr. 2019, doi: 10.3390/jcm8040462.

[6] S. Guo and Z. Yang, "Multi-channel-ResNet: An integration framework towards skin lesion analysis," *Informat. Med. Unlocked*, vol. 12, pp. 67–74, 2018, doi: 10.1016/j.imu.2018.06.006.

[7] W. Hua, T. Xiao, X. Jiang, Z. Liu, M. Wang, H. Zheng, and 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," *Biomed. Signal Process. Control*, vol. 58, Apr. 2020, Art. no. 101869, doi: 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, and P. Lambin, "Decision support systems in oncology," *JCO Clin. Cancer Informat.*, no. 3, pp. 1–9, Dec. 2019, doi: 10.1200/CCI.18.00001.

[9] C. Garling, "Andrew Ng: Why 'deep learning'is a mandate for humans, not just machines," *Wired*, May 2015. Accessed: Oct. 8, 2020. [Online]. Available: https://www.wired.com/brandlab/2015/05/andrew-ng-deep-learning-mandate-humans-not-just-machines/

[10] R. Sullivan, "Delivering affordable cancer care in high-income countries," *Lancet Oncol.*, vol. 12, no. 10, p. 48, 2011, doi: 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, and 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 Oncol.*, vol. 121, no. 3, pp. 459–467, Dec. 2016, doi: 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, and A. Dekker, "'Rapid learning health care in oncology'—An approach towards decision support systems enabling customised radiotherapy,'" *Radiotherapy Oncol.*, vol. 109, no. 1, pp. 159–164, Oct. 2013, doi: 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, and P. Lambin, "Systematic review of privacy-preserving distributed machine learning from federated databases in health care," *JCO Clin. Cancer Informat.*, vol. 4, pp. 184–200, Sep. 2020, doi: 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, and P. Lambin, "Infrastructure and distributed learning methodology for privacy-preserving multi-centric rapid learning health care: EuroCAT," *Clin. Transl. Radiat. Oncol.*, vol. 4, pp. 24–31, Jun. 2017, doi: 10.1016/j.ctro.2016.12.004.

[15] T. M. Deist *et al.*, "Machine learning algorithms for outcome prediction in (chemo)radiotherapy: An empirical comparison of classifiers," *Med. Phys.*, vol. 45, no. 7, pp. 3449–3459, Jul. 2018, doi: 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, and L. Wee, "Distributed radiomics as a signature validation study using the personal health train infrastructure," *Sci. Data*, vol. 6, no. 1, p. 218, Dec. 2019, doi: 10.1038/s41597-019-0241-0.

[17] M. J. Sheller, G. A. Reina, B. Edwards, J. Martin, and S. Bakas, "Multi-institutional deep learning modeling without sharing patient data: A feasibility study on brain tumor segmentation," in *Brainlesion: Glioma, Multiple Sclerosis, Stroke Traumatic Brain Injuries*, vol. 11383, A. Crimi, S. Bakas, H. Kuijf, F. Keyvan, M. Reyes, and T. van Walsum, Eds. Cham, Switzerland: Springer, 2019, pp. 92–104.

[18] C. Song, T. Ristenpart, and V. Shmatikov, "Machine learning models that remember too much," in *Proc. ACM SIGSAC Conf. Comput. Commun. Secur.*, Dallas, TX, USA, Oct. 2017, pp. 587–601, doi: 10.1145/3133956.3134077.

[19] X. Chen, J. Ji, C. Luo, W. Liao, and P. Li, "When machine learning meets blockchain: A decentralized, privacy-preserving and secure design," in *Proc. IEEE Int. Conf. Big Data (Big Data)*, Seattle, WA, USA, Dec. 2018, pp. 1178–1187, doi: 10.1109/BigData.2018.8622598.

[20] J. Weng, J. Weng, J. Zhang, M. Li, Y. Zhang, and W. Luo, "DeepChain: Auditable and privacy-preserving deep learning with blockchain-based incentive," *IEEE Trans. Depend. Sec. Comput.*, early access, Nov. 8, 2019, doi: 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, and P. Lambin, "Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach," *Nature Commun.*, vol. 5, no. 1, p. 4006, Sep. 2014, doi: 10.1038/ncomms5006.

[22] T. M. Mitchell, *Machine Learning*. New York, NY, USA: McGraw-Hill, 1997.

[23] K. Kourou, T. P. Exarchos, K. P. Exarchos, M. V. Karamouzis, and D. I. Fotiadis, "Machine learning applications in cancer prognosis and prediction," *Comput. Struct. Biotechnol. J.*, vol. 13, pp. 8–17, Jan. 2015, doi: 10.1016/j.csbj.2014.11.005.

[24] T. M. Deist *et al.*, "Distributed learning on 20 000+ lung cancer patients – the personal health train," *Radiotherapy Oncol.*, vol. 144, pp. 189–200, Mar. 2020, doi: 10.1016/j.radonc.2019.11.019.

[25] C. S. Wright, "Bitcoin: A peer-to-peer electronic cash system," White Paper. [Online]. Available: https://bitcoin.org/bitcoin.pdf

[26] P. Zhang, J. White, D. C. Schmidt, G. Lenz, and S. T. Rosenbloom, "FHIRChain: Applying blockchain to securely and scalably share clinical data," *Comput. Struct. Biotechnol. J.*, vol. 16, pp. 267–278, Jan. 2018, doi: 10.1016/j.csbj.2018.07.004.

[27] A. Havelange, M. Dumontier, B. Wouters, J. Linde, D. Townend, A. Riedl, and V. Urovi, "LUCE: A blockchain solution for monitoring data license accoUntability and CompliancE," 2019, *arXiv:1908.02287*. [Online]. Available: http://arxiv.org/abs/1908.02287

[28] V. Jaiman and V. Urovi, "A consent model for blockchain-based health data sharing platforms," *IEEE Access*, vol. 8, pp. 143734–143745, 2020, doi: 10.1109/ACCESS.2020.3014565.

[29] S. Vruddhula, "Application of on-dose identification and blockchain to prevent drug counterfeiting," *Pathogens Global Health*, vol. 112, no. 4, p. 161, May 2018, doi: 10.1080/20477724.2018.1503268.

[30] T.-T. Kuo, H.-E. Kim, and L. Ohno-Machado, "Blockchain distributed ledger technologies for biomedical and health care applications," *J. Amer. Med. Inform. Assoc.*, vol. 24, no. 6, pp. 1211–1220, Nov. 2017, doi: 10.1093/jamia/ocx068.

[31] H.-N. Dai, Z. Zheng, and Y. Zhang, "Blockchain for Internet of Things: A survey," *IEEE Internet Things J.*, vol. 6, no. 5, pp. 8076–8094, Oct. 2019, doi: 10.1109/JIOT.2019.2920987.

[32] C. D. Clack, V. A. Bakshi, and L. Braine, "Smart contract templates: Foundations, design landscape and research directions," 2016, *arXiv:1608.00771*. [Online]. Available: http://arxiv.org/abs/1608.00771

[33] Z. Zheng, S. Xie, H.-N. Dai, W. Chen, X. Chen, J. Weng, and M. Imran, "An overview on smart contracts: Challenges, advances and platforms," *Future Gener. Comput. Syst.*, vol. 105, pp. 475–491, Apr. 2020, doi: 10.1016/j.future.2019.12.019.

[34] 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, and P. Lambin, "Data from NSCLC-radiomics [data set]," *Cancer Imag. Arch.*, 2019, doi: 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, and F. Prior, "The cancer imaging archive (TCIA): Maintaining and operating a public information repository," *J. Digit. Imag.*, vol. 26, no. 6, pp. 1045–1057, Dec. 2013, doi: 10.1007/s10278-013-9622-7.

[36] R. A. Fisher, "The use of multiple measurements in taxonomic problems," *Ann. Eugenics*, vol. 7, no. 2, pp. 179–188, Sep. 1936, doi: 10.1111/j.1469-1809.1936.tb02137.x.

[37] D. G. Wood, "ETHEREUM: A secure decentralised generalised transaction ledger," ETHEREUM & ETHCORE, Tech. Rep. EIP-150 REVISION, p. 32. [Online]. Available: https://gavwood.com/paper.pdf

[38] Ethereum. *Remix*. Accessed: Aug. 14, 2020. [Online]. Available: https://remix.ethereum.org/

[39] E. Barker and A. Roginsky, "Transitioning the use of cryptographic algorithms and key lengths," Nat. Inst. Standards Technol., Gaithersburg, MD, USA, Tech. Rep. NIST SP 800-131Ar2, Mar. 2019, doi: 10.6028/NIST.SP.800-131Ar2.

[40] *Ganache*. Accessed: Aug. 14, 2020. [Online]. Available: https://www.trufflesuite.com/ganache

[41] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, Jun. 2016, pp. 770–778, doi: 10.1109/CVPR.2016.90.

[42] C. Li, J. Xu, Q. Liu, Y. Zhou, L. Mou, Z. Pu, Y. Xia, H. Zheng, and S. Wang, "Multi-view mammographic density classification by dilated and attention-guided residual learning," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, early access, Feb. 3, 2020, doi: 10.1109/TCBB.2020.2970713.

[43] E. R. DeLong, D. M. DeLong, and D. L. Clarke-Pearson, "Comparing the areas under two or more correlated receiver operating characteristic curves: A nonparametric approach," *Biometrics*, vol. 44, no. 3, p. 837, Sep. 1988, doi: 10.2307/2531595.

[44] X. Wu, J. Zhang, and F.-Y. Wang, "Stability-based generalization analysis of distributed learning algorithms for big data," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 31, no. 3, pp. 801–812, Mar. 2020, doi: 10.1109/TNNLS.2019.2910188.

[45] T.-T. Kuo, R. A. Gabriel, and L. Ohno-Machado, "Fair compute loads enabled by blockchain: Sharing models by alternating client and server roles," *J. Amer. Med. Inform. Assoc.*, vol. 26, no. 5, pp. 392–403, May 2019, doi: 10.1093/jamia/ocy180.

[46] T.-T. Kuo, J. Kim, and R. A. Gabriel, "Privacy-preserving model learning on a blockchain network-of-networks," *J. Amer. Med. Inform. Assoc.*, vol. 27, no. 3, pp. 343–354, Mar. 2020, doi: 10.1093/jamia/ocz214.

[47] S. Lugan, P. Desbordes, E. Brion, L. X. Ramos Tormo, A. Legay, and B. Macq, "Secure architectures implementing trusted coalitions for blockchained distributed learning (TCLearn)," *IEEE Access*, vol. 7, pp. 181789–181799, 2019, doi: 10.1109/ACCESS.2019.2959220.

[48] Q. Zhang, L. T. Yang, and Z. Chen, "Privacy preserving deep computation model on cloud for big data feature learning," *IEEE Trans. Comput.*, vol. 65, no. 5, pp. 1351–1362, May 2016, doi: 10.1109/TC.2015.2470255.

[49] R. Shokri and V. Shmatikov, "Privacy-preserving deep learning," in *Proc. 22nd ACM SIGSAC Conf. Comput. Commun. Secur. (CCS)*, Denver, CO, USA, 2015, pp. 1310–1321, doi: 10.1145/2810103.2813687.

[50] Y. Lu, X. Huang, Y. Dai, S. Maharjan, and Y. Zhang, "Blockchain and federated learning for privacy-preserved data sharing in industrial IoT," *IEEE Trans. Ind. Informat.*, vol. 16, no. 6, pp. 4177–4186, Jun. 2020, doi: 10.1109/TII.2019.2942190.

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

[52] J. Platt, "Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods," *Adv. Large Margin Classifiers*, vol. 10, pp. 61–74, Mar. 1999, doi: 10.1.1.41.1639.

[53] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," 2017, *arXiv:1706.04599*. [Online]. Available: https://arxiv.org/abs/1706.04599

FADILA ZERKA received the master's degree from the Technical University of Lodz, Poland, in 2014. She is currently pursuing the Ph.D. degree with Oncoradiomics, under the supervision of Prof. Philippe Lambin. She is a Computer Scientist. Her research is centered on privacy preserving distributed learning for optimal radiomics knowledge.

VISARA UROVI is currently an Associate Professor with the Institute of Data Science at Maastricht University. Her research focuses on blockchain-based solutions for data interoperability and exchange and standard-based privacy preserving distributed data exchange protocols. She has been involved in generic platform solutions for electronic health record data interoperability and exchange, and she was responsible for the definition of automated negotiation protocols in the ARGUGRID, FP7 EU project. She is the primary investigator of the PRIMA project, a blockchain-based solution focusing on health data exchange. She is currently working on the LUCE platform, an open-source blockchain platform for monitoring data licenses.

AKSHAYAA VAIDYANATHAN received the master's degree in cognitive psychology from the University College of Dublin. She is currently pursuing the Ph.D. degree with Oncoradiomics. Her research focuses on the application of AI in biomedical imaging.

SAMIR BARAKAT has 15 years of experience in data science, in clinical, academic, and commercial settings. Following the completion of his academic training, he joined a multi-disciplinary group at the universities of Sydney, Wollongong, and NSW applying his data mining and machine learning expertise to build, develop, and test models for clinical decision support systems using national and international data sources. He is mainly advising on research and development of DistriM in addition to supporting the continuous automated performance progression of RadiomiX.

RALPH T. H. LEIJENAAR is currently pursuing the Ph.D. degree with Maastricht University, with key expertise in, among others, advanced medical image analysis and machine learning (big data), more specifically in the field of radiomics (watch the animation). He is a Biomedical Engineer with Maastricht University. Being one of the pioneers of radiomics, he has deep expertise, experience, and understanding of this multidisciplinary field along with extensive clinical and pre-clinical research knowledge. He is consequently recognized as a preeminent author in radiomics literature (see Google Scholar). In 2015, he became involved in the valorization of research activities. He is co-founder and Chief Technology Officer of OncoRadiomics SA, Liège, Belgium.

SEAN WALSH is currently a Medical Physicist with key expertise in data science and a decade of experience in the fields of radiology and radiotherapy. He has a proven track record of managing patient data from multiple international cancer centers throughout the world computer assisted theragnostics network. He is primarily advising on clinical and pre-clinical research strategy, along with the data warehousing, distributed learning, network security and semantic web technology for the development of DistriM, while supporting the machine learning and signature progression of RadiomiX.

HANIF GABRANI-JUMA received the bachelor's degree in biomedical and electrical engineering and the master's degree in applied science from Carleton University, Ottawa, Canada. He is currently a Software Engineer working on applications in precision medicine and drug development with Oncoradiomics. He specializes in the development and validation of novel medical image analysis tools. His current interests include the efficient integration of computer-aided diagnosis tools into data processing pipelines, deriving clinical insights by leveraging large volumes of patient data, and synthetic disease modeling.

BENJAMIN MIRAGLIO is currently an Artificial Intelligence Scientist with Oncoradiomics. He has broad expertise in biology, information technology and computational sciences along with an extensive experience working as a computational biologist in the fields of genomics and toxicology, both in academy and industry. At Oncoradiomics, he adapts deep learning techniques in order to enhance the predictive capacities of RadiomiX.

HENRY C. WOODRUFF is currently the Lead Scientist with the D-Lab within the Department of Precision Medicine at Maastricht University. He gained his experience with image processing and analysis during the Ph.D. degree in astrophysics from Sydney University, in 2009. Since joining in 2017, he has pushed the forefront of radiomics, i.e., both the extraction as well as the machine learning for analysis and interpretation of multifactorial features from radiology images. Since 2018, he is an Appointed Deputy Head of The D-Lab. His expertise is in the application of machine learning methods and image processing methods on vast amounts of medical data in order to improve the lives of patients. For his work, he received a KWF high risk grant in 2018, and has published extensively in the field.

MICHEL DUMONTIER is currently a Distinguished Professor of data science with Maastricht University, and the Director of the Institute of Data Science. His research focuses on the development of computational methods for scalable integration and reproducible analysis of FAIR (Findable, Accessible, Interoperable and Reusable) data across scales - from molecules, tissues, organs, individuals, populations to the environment. His group combines semantic web technologies with effective indexing, machine learning and network analysis for drug discovery and personalized medicine. He is the scientific director for Bio2RDF, an open source project that provides Linked Data for the life sciences. He is a member of the Dutch Tech Center for Life Sciences, and participates in Elixir, a EU wide initiative to develop a distributed infrastructure for life-science information.

PHILIPPE LAMBIN received the Ph.D. degree in molecular biology. He is currently a Clinician, Radiation Oncologist, and pioneer in translational research with a focus on hypoxia and decision support systems. He is currently the Head of the Department of Precision Medicine (The D-Lab & M-Lab), Maastricht University. He is a ''ERC advanced & ERC PoC (2) grant laureate'' in 2016, 2018, and 2020, and he is coauthor of more than 500 peer reviewed scientific papers ($H_{GoogleScholar}$: 99), co-inventor of Distributed Learning, more than 20 patents (filed or submitted) and (co) promoter of more than 60 completed Ph.D.'s (two with cum laude). He is primarily advising on the scientific / Research and Development roadmap as well as lending his expertise to Oncoradiomics, MedC2, Convert Pharma, and LivingMed Biotech development strategy.