Encrypted federated learning for secure decentralized collaboration in cancer image analysis

One Sentence Summary:
Federated learning with homomorphic encryption enables multiple parties to securely co-train artificial intelligence models in pathology and radiology, reaching state-of-the-art performance with privacy guarantees.

Authors:
Daniel Truhn (1), Soroosh Tayebi Arasteh (1), Oliver Lester Saldanha (2,3), Gustav Müller-Franzes (1), Firas Khader (1), Philip Quirke (4), Nicholas P. West (4), Richard Gray (5), Gordon G. A. Hutchins (4), Jacqueline A. James (6,7,8), Maurice B. Loughrey (8,9,10), Manuel Salto-Tellez (6,7,8), Hermann Brenner (11, 12, 13), Alexander Brobeil (14,15), Tanwei Yuan (11, 17), Jenny Chang-Claude (17, 18), Michael Hoffmeister (11), Sebastian Foersch (19), Tianyu Han (20), Sebastian Keil (1), Maximilian Schulze-Hagen (1), Peter Isfort (1), Philipp Bruners (1), Georgios Kaisiss (21, 22, 23, 24), Christiane Kuhl (1), Sven Nebelung (1), Jakob Nikolas Kather (2,3,4,25)

Affiliations:
(1) Department of Diagnostic and Interventional Radiology, University Hospital RWTH Aachen, Aachen, Germany.
(2) Department of Medicine III, University Hospital RWTH Aachen, Aachen, Germany.
(3) Else Kroener Fresenius Center for Digital Health, Medical Faculty Carl Gustav Carus, Technical University Dresden, Dresden, Germany.
(4) Division of Pathology and Data Analytics, Leeds Institute of Medical Research at St James's, University of Leeds, Leeds, UK.
(5) Clinical Trial Service Unit, University of Oxford, Oxford, United Kingdom.
(6) Precision Medicine Centre of Excellence, Health Sciences Building, The Patrick G Johnston Centre for Cancer Research, Queen's University Belfast, United Kingdom.
(7) Regional Molecular Diagnostic Service, Belfast Health and Social Care Trust, Belfast, United Kingdom.
(8) The Patrick G Johnston Centre for Cancer Research, Queen's University Belfast, United Kingdom.
(9) Department of Cellular Pathology, Belfast Health and Social Care Trust, Belfast, United Kingdom.
(10) Centre for Public Health, Queen's University Belfast, United Kingdom.
(11) Division of Clinical Epidemiology and Aging Research, German Cancer Research Center (DKFZ), Heidelberg, Germany.
(12) Division of Preventive Oncology, German Cancer Research Center (DKFZ) and National Center for Tumor Diseases (NCT), Heidelberg, Germany.
(13) German Cancer Consortium (DKTK), German Cancer Research Center (DKFZ), Heidelberg, Germany.
(14) Institute of Pathology, University Hospital Heidelberg, Heidelberg, Germany.
(15) Tissue Bank, National Center for Tumor Diseases (NCT), University Hospital Heidelberg, Heidelberg, Germany.
(16) Medical Faculty Heidelberg, Heidelberg University, Heidelberg.
(17) Cancer Epidemiology Group, University Cancer Center Hamburg, University Medical Center Hamburg-Eppendorf, Hamburg, Germany.
(18) Division of Cancer Epidemiology, German Cancer Research Center (DKFZ), Heidelberg, Germany.
Abstract:

Artificial Intelligence (AI) has a multitude of applications in cancer research and oncology. However, the training of AI systems is impeded by the limited availability of large datasets due to data protection requirements and other regulatory obstacles. Federated and swarm learning represent possible solutions to this problem by collaboratively training AI models while avoiding data transfer. However, in these decentralized methods, weight updates are still transferred to the aggregation server for merging the models. This leaves the possibility for a breach of data privacy, for example by model inversion or membership inference attacks by untrusted servers. Homomorphically encrypted federated learning (HEFL) is a solution to this problem because only encrypted weights are transferred, and model updates are performed in the encrypted space. Here, we demonstrate the first successful implementation of HEFL in a range of clinically relevant tasks in cancer image analysis on multicentric datasets in radiology and histopathology. We show that HEFL enables the training of AI models which outperform locally trained models and perform on par with models which are centrally trained. In the future, HEFL can enable multiple institutions to co-train AI models without forsaking data governance and without ever transmitting any decryptable data to untrusted servers.
Introduction

Artificial intelligence (AI) and machine learning techniques are transforming cancer imaging and cancer research and will have a profound impact on the practice of medicine\textsuperscript{1–4}. They can automate manual tasks in medical image analysis and can be used to extract hidden information from routinely available clinical image data, beyond what is visible to the human eye\textsuperscript{5,6}. AI models have been used for the detection and diagnosis of cancer, subtype classification, and optimization of cancer treatments. In particular, deep neural networks have been trained to analyze radiology images and digitized pathology slides for numerous different cancer types. For example, AI models can now detect mammographic lesions with expert-level performance\textsuperscript{7}. Similarly, AI models predict molecular biomarkers for treatment selection directly from routine pathology slides of solid tumors\textsuperscript{8–13}.

However, the training of AI models is infamously data hungry and requires large amounts of annotated training data. While this data may already exist, in most cases it is scattered among multiple centers. Collecting this data at a central site is hindered by obstacles which are often insurmountable in practice, most notably issues with data privacy and data governance. The data governance problem has been addressed by collaborative learning protocols such as federated learning (FL)\textsuperscript{14,15} in which an AI model is trained on separate sites and in which not data, but only the learned model weights are shared. This facilitates collaboration between multiple parties, but still poses significant risks for breach of patient privacy. The weight updates communicated to the central FL server contain information about the data that can be extracted to reconstruct sensitive patient information\textsuperscript{16}. This can be exploited through privacy attacks such as model inversion\textsuperscript{17–19}, in which a malicious server eavesdropper captures the weight updates and attempts to recover the private dataset used to train the model or reveal other private attributes. One measure to protect against privacy breaches is Differential Privacy in which deliberate noise is added to the training updates by each site \textsuperscript{18}. However, while this paradigm protects private information, it comes at a utility tradeoff and can lead to less performant AI models as demonstrated recently\textsuperscript{20}. In our study we employ Homomorphic encryption (HE). HE can protect against a malicious server eavesdropper while maintaining AI model performance by encrypting the weight updates before sending them to the
central server. The central server then performs the weight aggregation on the encrypted values and
the encrypted updated weights are sent back to the clients for decryption and incorporation into their
models. Importantly, since the central server does not have access to the decryption key, it cannot
infer any information about which calculations have been done at individual peer locations and thus
cannot extract sensitive private information. In other words, all handling of the model parameters
happens in the encrypted space, making homomorphic encryption an optimal tool for low-trust
environments and handling of personal health data.

In this study, we examined how HEFL can be leveraged for training of competitive AI models for
cancer diagnosis and detection of cancer biomarkers in radiology and pathology images. To this end,
we assumed the following threat model: A mutually trusting confederation of data owners wishes to
collaboratively train a model on their joint data, but neither wants to relinquish data governance. For
conducting the training, the confederation makes use of an untrusted aggregation server, which we
assume to honestly participate in the protocol (i.e., faithfully conduct the aggregation procedure), but
attempt to extract all available information from the weight updates sent to it by the other participants
(honest but curious threat model). We evaluated the training of AI models in three retrospective
multicentric settings: 1) AI models are trained with local data only 2) AI models are trained with
conventional federated learning whereby no additional measure of protection against privacy-centred
attacks on the updates is undertaken and 3) AI models are trained with HEFL in a decentralized,
secure and privacy-preserving manner, whereby the individual participants encrypt their weight
update before transmitting it to the server. We hypothesized that the collective and secure training of
AI models reaches better accuracy than training of local models and is associated with minimal risk of
privacy leakage as compared to conventional FL while keeping the cost of additional training time
low.
Results

**HEFL guarantees data privacy compared to conventional federated learning in the untrusted central server setting**

When multiple institutions collaborate in a conventional federated learning scheme, weight updates are calculated locally and are sent to a central server to be aggregated. When unencrypted weight updates are transmitted, we demonstrate that the untrusted central server can reconstruct the training images from the weight updates in a model inversion attack. In this setting we train a neural network for the detection of malignant lesions on brain MRI examinations from the brain tumor segmentation (BraTS) dataset. We employ a realistic setting in which data is contributed by five different institutions and in which each institution performs separate weight updates only on their data. We then perform a gradient inversion attack following the approach by Zhao et al. We demonstrate that the original training images can be reconstructed after only 120 iterations - notably, before training of the underlying neural network objective has converged, see Figure 1. This poses a serious threat and renders the whole concept of conventional federated learning vulnerable to privacy-focused attacks. To showcase that homomorphic encryption can be used to counter these attacks and to salvage patient privacy, we repeat the training procedure, but employ homomorphic encryption in which the central server only has access to the encrypted weight updates and the key is kept private by the peers. Following the same approach - no identifiable information can be extracted from the weight updates, even after eavesdropping on the weight updates for 40,000 iterations.

**Secure training does not affect performance of oncological AI models**

We trained AI models for tasks in oncology spanning both radiological and histopathological use-cases, see Figure 2. Each model was trained in three settings: a) AI models are trained with local data only b) AI models are trained with conventional federated learning in a decentralized manner c) AI models are trained with HEFL in a decentralized, secure and privacy-preserving manner. While
approach a) is immune to privacy leaks, it results in training on only a limited subset of the possible
data pool. Approach b) makes full use of the data but is prone to privacy leaks through the
aforementioned attack by the untrusted aggregator. Only approach c) combines both training on full
data and guarantees patient privacy. Moreover, as the HE scheme utilized in our study is endowed
with a correctness guarantee (i.e. the values of the decrypted updates are guaranteed to be identical
up to numerical precision to their plain-text counterparts), this setting does not suffer from an
accuracy penalty compared to non-private training. We test the performance of each paradigm for AI
models for the segmentation of glioblastoma on magnetic resonance images (MRI) and for the
detection of microsatellite instability in histopathological whole slide images (WSI) of colorectal
cancer patients.

Segmentation of glioblastoma on MRI

The BraTS training dataset comprises 369 MRI examinations of 369 patients which have been
acquired at seventeen different clinical centers. We partitioned the data along the information where
the images had been acquired into five groups and trained a U-Net architecture to segment the tumor
volume. All models were tested on an external test set from a separate institution provided by the
BraTS organizers (n=125) and employed the dice similarity score as a measure of performance. All
five locally trained AI models performed inferior in terms of the dice score both to the models trained
with FL and with HEFL. Notably, no performance drop was seen in the model trained with HEFL as
compared to the model trained with conventional FL, cf. Table 1.

Prediction of genetic biomarkers in colorectal cancer patients from pathology images

In an analogous setting to the radiological use-case, we tested whether HEFL performs equal to
conventional FL and superior to locally trained models in the benchmarking task of predicting a
molecular biomarker in colorectal cancer from pathology images: microsatellite instability
(MSI)/mismatch repair deficiency (dMMR), which qualifies metastatic patients to receive cancer immunotherapy.

We performed the evaluation on independent test sets never seen during training: the clinical trial cohort QUASAR (n = 1,774 patients from the United Kingdom) and the population-based cohort YCR BCIP (Yorkshire Cancer Research Bowel Cancer Improvement Programme, n=889 patients). We trained three models on the Epi700 data (United Kingdom, n=607), the DACHS data (Germany, n=2,039) and the TCGA data (USA, n=426) respectively. Subsequently, we trained one model each in the federated learning setup including all three datasets without and with homomorphic encryption. Training with HEFL was superior to training just with local data and non-inferior to training with FL both for testing on the YCR cohort and for testing on the QUASAR cohort. Both FL and HEFL performed on the same level with no detectable difference, cf. Table 2.

**Secure training is time-efficient**

A notable drawback of homomorphic encryption is its computational overhead. In our study, we eschewed this drawback by encrypting not the entire training process, but only the privacy-critical weight aggregation step, which is performed by a (potentially untrusted third party), thus enabling substantial computational savings. To determine the effect of our scheme on training time compared to FL without encryption, we conducted the following experiments on a typical hardware setup used in machine learning. As a side note, de- and encryption as well as weight aggregation is usually conducted on the central processing unit (CPU), while backpropagation during training of the networks depends on the graphics processing unit (GPU).

We found that the time required for encryption was almost negligible compared to the time required to perform the backpropagation steps and the application of weight updates: for the radiological use-case described above, less than 1% of computational time was spent on decryption, encryption and homomorphic aggregation of the weights (Figure 3d). For the histopathological use-case, less than 5% of time was used for decryption and encryption (which happens at edge) and homomorphic
aggregation of the weights (which happens at the central server, Figure 3b). This difference is due to the different network architectures and different number of parameters: the histopathological use-case employs a fixed backbone feature extractor and thus has fewer parameters to optimize.

Encryption and decryption scales approximately linear with the number of weights to be updated, while neural network training complexity scales more than linearly in our setup. Thus, more complex networks, such as the one used to segment brain tumors invest more computational resources in the backpropagation algorithm relative to the encryption algorithm. This is encouraging, since the relationship between training time and aggregation time is in favor of more complex networks that are usually employed when working with big data.

Discussion

AI has an indisputable potential in the field of oncology and AI models are currently reaching a stage in which they can improve patient care and render medical processes more efficient. However, this improvement critically depends on the availability of sufficiently large, curated, and representative training data. Currently, most research groups and industry have limited and only local data access. To train useful and generalizable AI models, stakeholders need to be able to collaborate on a large scale without jeopardizing patient privacy. Only through such multi-institutional collaboration can robust AI models be trained that adequately capture the entire human population and that make the transition from bench to bedside. Federated learning was initially proposed as a technical solution for privacy-preserving distributed AI. FL enables joint training of AI models by multiple partners who do not share their data with each other and has been demonstrated to facilitate the training of AI models on big data. Similarly, swarm learning (SL) utilizes a network of nodes to jointly train a model on distributed data and to aggregate model weights without a central instance. However, FL and SL have an important shortcoming: during training, weight updates must be shared and information about the underlying data can be extracted from...
these weight updates as shown in our study. Such techniques should thus not be considered privacy techniques, but techniques for preserving data governance\textsuperscript{32}. Since medical data is highly sensitive and since data privacy laws forbid the use of data in such environments, where private data can be extracted, this critically limits the applicability of collaborative learning schemes and prevents the development of powerful AI models in cancer diagnosis and treatment.

This shortcoming can be remedied by employing techniques which guarantee privacy to data owners. The only technique to guarantee privacy in a data release process is Differential Privacy\textsuperscript{33}. Hence, when sharing the model with untrusted third parties, such a technique would have to be employed to constrain the success of attacks against patient privacy. We operate under a slightly different threat model. As all participants of the federated learning workflow described above are mutually trusting, are not intending to publish the model to the outside world and all receive an identical copy of the final model, we need only protect against an attack by the (untrusted aggregation server). Our homomorphic encryption scheme protects the weights during this critical aggregation step: local sites encrypt their weight updates before sending them out and keep the decryption key private. The entity which receives the weight updates from all sites and which performs the weight aggregation in the encrypted space thus has no access to the underlying data and no sensitive data can be extracted by design. Our technique has two notable benefits: it sidesteps the computational overhead of having to train the entire model in the encrypted space using HE. In principle, it would also be possible to use HE on all levels of the training process - i.e., also during backpropagation. However, with concurrently available computational resources, this has proven to be prohibitively computationally expensive and is not yet in reach\textsuperscript{34}. Furthermore, as long as all data stays on site - as is the case in our FL setup - there is no need to encrypt the backpropagation procedure: potential eavesdroppers do not have access to that part of the training procedure as it is done behind secure firewalls. By restricting the fully homomorphic encryption to the most critical part of FL - the weight aggregation - we show that additional computational overload is almost negligible. Moreover, our technique allows us to avoid
the privacy-utility trade-offs of employing Differential Privacy for training, in which training with
Differential Privacy can lead to less-performant AI models\textsuperscript{20}. We note that the utilization of Differential Privacy would be mandatory in threat models different from ours, especially if the final model was designed to be shared with untrusted third parties.

A similar scheme to ours was demonstrated by Kaissis et al. in a proof-of-concept study for classifying pneumonia on chest radiographs by using secure multi-party computation through additive secret sharing\textsuperscript{1634}. With our study, we are the first to comprehensively assess fully homomorphic encryption in cancer diagnosis on large multi-centric databases spanning both radiology and histopathology.

Our study demonstrates that AI models for oncological image-processing can be trained securely on multi-institutional data without compromising patient privacy. This will facilitate collaboration between researchers and industry alike, ultimately leading to the development of advanced and clinically useful AI models. We show that implementing the FL scheme together with homomorphic encryption comes with minimal additional code complexity and can be performed with our publicly available code.

A technical limitation of our study is that we performed all experiments within one institutional network. However, by distributing the datasets to different computing entities and keeping them strictly separate, we simulated the setting in which multiple institutions - each with their own network - perform FL realistically. We assumed a constant network communication cost in our experiments. In realistic settings, communication overhead can be unpredictable, as it depends on more factors than network size (such as concurrent traffic or physical distance of the sites). We thus chose to exclude this factor, believing it to only represent a minor limitation. We note that homomorphically encrypted weights cannot be efficiently reduced in size by compression, however this limitation is negligible compared to the requirement to encode them as 64-bit data types for transmission over HTTP. Moreover, as all parties are mutually trusting and receive an identical copy of the fully trained model.
at the end of training, we utilized the same key pair to encrypt the weights on all participating nodes, thus avoiding the technical challenge of key distribution.

Further improvements to the FL process are possible: with increasing peer numbers who participate in the FL setup, participation of a bounded number of malicious participants who try to corrupt the training process by delivering adversarial weight updates is possible, whereas we regarded all participants as either fully trusted or honest but curious. It has been shown that regular FL fails to converge in the presence of faulty and malicious clients\textsuperscript{35}. Measures to counter these attacks are available and can be integrated in federated learning schemes should the need arise \textsuperscript{36}.

In conclusion, our study provides a blueprint for the secure and privacy-preserving multi-institutional training of oncological AI models and solves an urgent need, since it is becoming increasingly clear that differences in race and gender affect disease risk among individuals and that existing datasets at local institutions are insufficient to account for these effects.

**Methods**

**Ethics statement**

This study was carried out in accordance with the Declaration of Helsinki. This study is a retrospective analysis of publicly available anonymized MRI examinations and of anonymized histopathological tissue samples from multiple cohorts of cancer patients. Collection and anonymization of patients in all cohorts took place in each contributing center. Approval by the local ethics committee at each contributing center was given if applicable (QUASAR: North East – York Research Ethics Committee; YCR: Ethical approval was not required, because screening was recommended in all patients diagnosed with CRC. Testing was considered part of the ‘standard of care’ clinical pathway; Epi700: Northern Ireland Biobank (NIB13/0069, NIB13/0087, NIB13/0088 and NIB15/0168), DACHS: Ethics committee of the Medical Faculty, University of Heidelberg). Approval
of the ethics committee at the University Hospital of Aachen was given for the retrospective analysis of anonymized image data under reference number “Ethikkommission EK 028/19”.

**Patient cohorts**

MRI data for the BraTS patient collective contains brain MRI scans of 341 patients collected from 17 imaging centers and additional 28 patients for whom the imaging centers were not specified by the data provider. During federated learning we allocated the patients to five data clusters simulating the situation in which a regional hospital’s image database contains MRI data of multiple imaging centers. This situation is typical in real-world scenarios where patients are referred for surgery and bring their image data that had been acquired at an external institution before. The allocation of patients is detailed in supplemental table 1. All MRI examinations contained pre- and post-contrast T1-weighted sequences, T2-weighted sequences and fluid attenuation inversion-recovery sequences (FLAIR). All sequences were acquired in axial orientation. All the imaging datasets have been segmented manually, by one to four raters, following the same annotation protocol, and their annotations were approved by experienced neuro-radiologists.

For the histopathological data we collected digital whole slide images (WSI) of H&E-stained slides of human colorectal cancer (CRC) from five patient cohorts, three of which were used as training cohorts and two of which were used as test cohorts following the division of data in a previous study. The training cohorts are representative of real-world clinical settings. First, the Northern Ireland Epi700 (n=661) cohort study contained data of patients with stage II and III colon cancer. This data was provided by the Northern Ireland Biobank (application NIB20-0346). Second, the “Darmkrebs: Chancen der Verhütung durch Screening” study (DACHS, n=2448) is a large population-based case-control study. This data includes samples of CRC patients at any disease stage. This data was collected from over 20 hospitals in Germany. Data collection was coordinated by the German Cancer Research Center (DKFZ, Heidelberg, Germany) and supported by the NCT tissue bank at the National Center for Tumor Diseases and the Institute of Pathology at the University of Heidelberg. Third, “The Cancer Genome Atlas” (TCGA) CRC cohort (n=632) is a large
collection of tissue specimens from multiple populations across different countries, but largely from the United States of America (USA).\textsuperscript{42}

We employed two separate test cohorts: The “Quick and Simple and Reliable” (QUASAR) cohort was derived from a clinical trial of adjuvant therapy containing 2206 WSI, which aimed to determine survival benefit from adjuvant chemotherapy in CRC patients from the United Kingdom (UK)\textsuperscript{43,44}. The second test cohort used data from the Yorkshire Cancer Research Bowel Cancer Improvement Programme\textsuperscript{45} (YCR-BCIP) cohort (n=889). This was a population-based study collected in the Yorkshire Region in the UK. For all cohorts, microsatellite instability (MSI) / mismatch repair deficiency (dMMR)\textsuperscript{46} data were acquired.

The distribution of tumor stages in TCGA, DACHS and YCR-BCIP is comparable, see supplemental table 2. In QUASAR, stage III tumors are overrepresented due to the fact that adjuvant therapy is mainly performed in intermediate stage tumors. Therefore, following previous work\textsuperscript{25}, we used YCR-BCIP and QUASAR as test cohorts to investigate the robustness of the AI models both on a general population and on a clinical trial population. Importantly, neither in the MRI data nor in the histopathological data, there was any overlap between training and test cohorts.

**Deep learning training procedure**

The hardware used in our experiments were Intel CPUs with 18 cores and 32 GB RAM and Nvidia RTX 6000 GPUs with 24 GB memory.

**MRI data**

All of the 3D volumes were cropped around the brain to lower the computational costs and standardize the field of view. As intensity distributions vary across magnetic resonance images, intensity normalization is crucial. Therefore, we clipped the intensity values above the 99 percentile of the image, then subtracted the minimum value of the result from voxel values and divided the shifted image by the maximum value of the image. We performed data augmentation during training by
applying random cropping of patches of $128 \times 128 \times 128$ from each original volume around its center. Additionally, we applied medio-lateral and cranio-caudal flipping with a probability of 0.4. Intensity was randomly rescaled according to a power-law $I_{\text{new}} = g \cdot I^\gamma$ with gain $g$ and the exponent $\gamma$ randomly selected between 0.8 - 1.2 from a uniform distribution. White Gaussian noise with zero mean and a standard deviation of 0.03 was added to each sequence of the multi modal MRI data.

A modified 4-level 3D U-Net \cite{48,49} was utilized for segmentation of brain tumors. In the contraction path, each layer contained two $3 \times 3 \times 3$ convolutions, each followed by a rectified linear unit (ReLU), a batch normalization (BN) and then a $2 \times 2 \times 2$ max pooling with strides of two in each dimension. The output channel number was doubled after each level in the contraction path, and it was 48 at the end of the level one. In the expansion path, each layer consisted of a nearest neighbor up-sampling of $2 \times 2 \times 2$ in each dimension, followed by two $3 \times 3 \times 3$ convolutions each followed by a ReLU and BN. The output channel number was halved after each level in the expansion path. In the last layer, a $1 \times 1 \times 1$ convolution, which reduced the number of output channels to 3, followed by a SoftMax layer, was used for the per-voxel final classification.

The model was optimized using the Adam optimizer with a learning rate of $10^{-4}$. To be consistent in our comparison scenarios, all the weight and bias parameters of all the different models were initialized using the He initialization scheme \cite{50}. As a loss function, we chose the Dice loss tailored to the BraTS data needs \cite{51}. To minimize the overhead and make maximum use of the graphics processing unit memory, we utilized large input tiles over a large batch size and reduced the batch to a single 3D image \cite{49} with 4 channels, each channel being one of the MR modalities. Hence, the batch normalization acted like instance normalization in our implementation. The network contained a total of 5,670,579 trainable parameters.

**Histopathological data**

For prediction of molecular features from image data, we based our analysis on a well established weakly-supervised end-to-end prediction pipeline, which was described and evaluated in a recent benchmark study. \cite{52}. As a preprocessing step, the original gigapixel WSIs were tessellated into
patches of size \((512 \times 512 \times 3)\) pixels and were color-normalized with the Macenko method.\(^{53}\)

Blurry patches and patches with no tissue were removed from the data set using canny edge detection \(^{52}\). Following that approach, we obtained a normalized edge image using the “canny” method in Python’s OpenCV package and then removed all tiles with a mean value below a threshold of 4. A pre-trained ResNet18 was used to extract a \((512 \times 1)\) feature vector from 150 randomly selected patches for each patient \(^9\). Before training, the number of tiles in each class were equalized by random undersampling until all classes had the same number of tiles, as described before \(^9,12\).

Feature vectors served as input to a fully connected classification network and the patient-wise MSI label was used to label every single tile derived from that patient. The fully connected classifier network comprised four layers with \((512 \times 256), (256 \times 256), (256 \times 128)\) and \((128 \times 2)\) connections with a ReLU activation function.

**Code availability**

Our source code for secure federated learning using homomorphic encryption is publicly available at [https://github.com/tayebiarasteh/federated_HE](https://github.com/tayebiarasteh/federated_HE) and will be permanently archived on zenodo upon publishing of the paper. All source codes for training and evaluation of the deep neural networks, MR image analysis and preprocessing, 3D data augmentation, and gradient inversion attack are available at [https://github.com/tayebiarasteh/federated_HE](https://github.com/tayebiarasteh/federated_HE). All source code for the histological image analysis is available at [https://github.com/KatherLab/HIA](https://github.com/KatherLab/HIA) and all source code for histological image preprocessing is available at [https://github.com/KatherLab/preProcessing](https://github.com/KatherLab/preProcessing). All code for the experiments was developed in Python 3.8 using the PyTorch 1.4 framework. The secure federated learning process was developed using PySyft 0.2.9.

**Data availability**

The data that support the findings of this study are in part publicly available, in part proprietary datasets provided under collaboration agreements. Data from the BraTS collective is publicly
available under https://www.med.upenn.edu/cbica/brats2020/data.html. Data (including histological images) from the TCGA database are available at https://portal.gdc.cancer.gov/. All molecular data for patients in the TCGA cohorts are available at https://cbioportal.org. Data access for the Northern Ireland Biobank can be requested at http://www.nibiobank.org/for-researchers. All other data can be requested from the respective study groups who independently manage data access for their study cohorts.
References:


50. He, K., Zhang, X., Ren, S. & Sun, J. Delving deep into rectifiers: Surpassing human-level

Neural Networks: A BraTS 2020 Challenge Solution. in Brainlesion: Glioma, Multiple Sclerosis,

52. Laleh, N. G. et al. Benchmarking artificial intelligence methods for end-to-end computational

Additional information

Acknowledgements

The authors are grateful for the support by NVIDIA who provided counsel and supported our group with two RTX6000 GPUs. We additionally acknowledge support by the tissue bank of the National Center for Tumor Diseases (NCT) at the Institute of Pathology at University Hospital Heidelberg, Heidelberg, Germany, for providing access to the biobank data.

Funding sources

JNK is supported by the German Federal Ministry of Health (DEEP LIVER, ZMVI1-2520DAT111) and the Max-Eder-Programme of the German Cancer Aid (grant #70113864). The DACHS study (H.B., J.C.-C. and M.H.) was supported by the German Research Council (BR 1704/6-1, BR 1704/6-3, BR 1704/6-4, CH 117/1-1, HO 5117/2-1, HO 5117/2-2, HE 5998/2-1, HE 5998/2-2, KL 2354/3-1, KL 2354/3-2, RO 2270/8-1, RO 2270/8-2, BR 1704/17-1 and BR 1704/17-2), the Interdisciplinary Research Program of the National Center for Tumor Diseases (NCT; Germany) and the German Federal Ministry of Education and Research (01KH0404, 01ER0814, 01ER0815, 01ER1505A and 01ER1505B). The Epi700 creation was enabled by funding from Cancer Research UK (C37703/A15333 and C50104/A17592) and a Northern Ireland HSC R&D Doctoral Research Fellowship (EAT/4905/13). P.Q. and N.P.W. are supported by Yorkshire Cancer Research Programme grants L386 (QUASAR series) and L394 (YCR BCIP series). P.Q. is a National Institute of Health Research senior investigator. J.A.J. has received funds from Health and Social Care Research and Development (HSC R&D) Division of the Public Health Agency in Northern Ireland (R4528CNR and R4732CNR) and the Friends of the Cancer Centre (R2641CNR) for development of the Northern Ireland Biobank.
**Author contributions**

JNK, DT and STA designed the study; STA, OLS, DT and JNK developed the software; STA performed the experiments; STA and DT and JNK analyzed the data; STA performed statistical analyses; all authors provided clinical expertise and contributed to the interpretation of the results. STA, DT and JNK wrote the manuscript, all authors corrected the manuscript and collectively made the decision to submit for publication.

**Competing interests**

The Authors declare no competing financial or non-financial interests. For transparency, we provide the following information: JNK declares consulting services for Owkin, France and Panakeia, UK. PQ and NW declare research funding from Roche and PQ consulting and speaker services for Roche. MST has recently received honoraria for advisory work in relation to the following companies: Incyte, MindPeak, MSD, BMS and Sonrai; these are all unrelated to this work. No other potential conflicts of interest are reported by any of the authors. The authors received advice from NVIDIA when performing this study, but NVIDIA did not have any role in study design, conducting the experiments, interpretation of the results or decision to submit for publication.
Fig. 1: Schematic of FL and HEFL and associated Information extraction attacks. (A) In FL, each site trains on their own data and weight updates are transmitted to the central server for aggregation. (B) In HEFL, the weight updates are encrypted and the server only has access to the encrypted values. While FL allows the server to extract patient sensitive information by reconstructing the images from the weights through gradient inversion.
attacks and eavesdropping on the weight updates (C), this information remains protected in HEFL and images can not be reconstructed (D). Experiments were performed on 2D slices including native T1-weighted sequences in the top row, post-contrast T1-weighted sequences in the second row, T2-weighted sequences in the third row and fluid attenuated inversion recovery sequences in the bottom row.
Fig. 2: Schematic of the Deep Learning workflow. (A) Histology images are first tessellated. Features are then extracted by a feature extractor network (fixed) and a multi-layer perceptron is trained to predict MSI status. (B) The MRI examination is normalized and rescaled to a standard resolution of 128x128x128. All four three-dimensional sequences are then fed into a UNet architecture that is trained to predict tumor segmentation outlines.
Fig. 3: Results of training on local data only vs. training using FL and HEFL. Training neural networks on single-site datasets results in inferior performance as compared to FL and HEFL. A neural network was trained to detect MSI on data from the Epi700, the DACHS and the TCGA cohorts respectively as well as on all three datasets using FL and HEFL. The resulting networks were then tested on the QUASAR (A) and the YCR BIP (B) cohorts demonstrating superior performance of FL and HEFL. Similarly, tumor segmentation in MRI data was trained on data from five different sites as well as on all data using FL and HEFL. The resulting neural networks were then tested on an
independent held-out test set and demonstrated improved performance (C). Computational overhead was almost negligible (red: overhead for FL, yellow: additional overhead for encryption) as compared to training time needed for backpropagation (blue), (D).
### Tables

<table>
<thead>
<tr>
<th>Dice Score (%)</th>
<th>Site 1</th>
<th>Site 2</th>
<th>Site 3</th>
<th>Site 4</th>
<th>Site 5</th>
<th>FL</th>
<th>HEFL</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>66.15 ± 29.56</td>
<td>78.43 ± 21.58</td>
<td>76.58 ± 22.89</td>
<td>77.63 ± 19.86</td>
<td>76.54 ± 22.57</td>
<td>81.71 ± 18.89</td>
<td>80.32 ± 19.40</td>
</tr>
<tr>
<td>(p=0.001)</td>
<td>(p=0.101)</td>
<td>(p=0.021)</td>
<td>(p=0.021)</td>
<td>(p=0.003)</td>
<td>(p=0.091)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Table 1:** Performance of the five radiological AI models that were trained on local data only (site 1-5) and of the AI model that was trained with federated learning (FL) and with additional homomorphic encryption (HEFL). P-values are given for the comparison to HEFL.

<table>
<thead>
<tr>
<th></th>
<th>Train on Epi700</th>
<th>Train on DACHS</th>
<th>Train on TCGA</th>
<th>FL</th>
<th>HEFL</th>
</tr>
</thead>
<tbody>
<tr>
<td>Testing on QUASAR</td>
<td>74.66 ± 1.50</td>
<td>70.38 ± 1.22</td>
<td>70.94 ± 1.68</td>
<td>78.52 ± 1.34</td>
<td>79.54 ± 1.45</td>
</tr>
<tr>
<td>(p=0.008)</td>
<td>(p=0.001)</td>
<td>(p=0.001)</td>
<td>(p=0.289)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Testing on YCR BCIP</td>
<td>77.13 ± 1.74</td>
<td>82.46 ± 2.08</td>
<td>78.83 ± 1.67</td>
<td>85.42 ± 1.63</td>
<td>86.77 ± 1.65</td>
</tr>
<tr>
<td>(p=0.001)</td>
<td>(p=0.054)</td>
<td>(p=0.001)</td>
<td>(p=0.270)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Table 2:** Area under the receiver-operator-curve for the histopathological AI models that were trained for MSI detection on the Epi700, DACHS and TCGA datasets respectively and tested on the independent QUASAR and YCR BCIP cohorts. P-values are given for the comparison to HEFL.
### Supplemental Information

<table>
<thead>
<tr>
<th>Data source</th>
<th>number of examinations</th>
<th>Allocation to Federated Learning Center #</th>
</tr>
</thead>
<tbody>
<tr>
<td>BraTS center 1</td>
<td>129</td>
<td>1</td>
</tr>
<tr>
<td>BraTS center 2</td>
<td>14</td>
<td>2</td>
</tr>
<tr>
<td>BraTS center 3</td>
<td>5</td>
<td>2</td>
</tr>
<tr>
<td>BraTS center 4</td>
<td>9</td>
<td>2</td>
</tr>
<tr>
<td>BraTS center 5</td>
<td>22</td>
<td>2</td>
</tr>
<tr>
<td>BraTS center 6</td>
<td>34</td>
<td>3</td>
</tr>
<tr>
<td>BraTS center 7</td>
<td>12</td>
<td>3</td>
</tr>
<tr>
<td>BraTS center 8</td>
<td>8</td>
<td>3</td>
</tr>
<tr>
<td>BraTS center 9</td>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td>BraTS center 10</td>
<td>8</td>
<td>4</td>
</tr>
<tr>
<td>BraTS center 11</td>
<td>6</td>
<td>4</td>
</tr>
<tr>
<td>BraTS center 12</td>
<td>9</td>
<td>4</td>
</tr>
<tr>
<td>BraTS center 13</td>
<td>27</td>
<td>4</td>
</tr>
<tr>
<td>BraTS center 14</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>BraTS center 15</td>
<td>11</td>
<td>4</td>
</tr>
<tr>
<td>BraTS center 16</td>
<td>34</td>
<td>5</td>
</tr>
<tr>
<td>BraTS center 17</td>
<td>6</td>
<td>5</td>
</tr>
<tr>
<td>BraTS unknown group</td>
<td>28</td>
<td>5</td>
</tr>
<tr>
<td>Sum</td>
<td>369</td>
<td></td>
</tr>
</tbody>
</table>

**Table S1:** Allocation of patients in the BraTS collective to five data clusters used in the federated learning setup simulating the situation in which five regional hospitals' image databases contain a multitude of examinations from different scanners.
<table>
<thead>
<tr>
<th></th>
<th>TCGA</th>
<th>DACHS</th>
<th>Epi700</th>
<th>QUASAR</th>
<th>YCR-BCIP</th>
</tr>
</thead>
<tbody>
<tr>
<td>N Patients</td>
<td>632</td>
<td>2448</td>
<td>661</td>
<td>2190</td>
<td>889</td>
</tr>
<tr>
<td>Age (median ± IQR)</td>
<td>68 ± 18</td>
<td>69 ± 14</td>
<td>72 ± 14</td>
<td>63 ± 12</td>
<td>71 ± 15</td>
</tr>
<tr>
<td>Gender: Male</td>
<td>322 (50.9%)</td>
<td>1436 (58.7%)</td>
<td>358 (54.2%)</td>
<td>1334 (60.9%)</td>
<td>494 (55.6%)</td>
</tr>
<tr>
<td>Gender: Female</td>
<td>292 (46.2%)</td>
<td>1012 (41.3%)</td>
<td>303 (45.8%)</td>
<td>848 (38.7%)</td>
<td>395 (44.4%)</td>
</tr>
<tr>
<td>Gender: Unknown</td>
<td>18 (2.85%)</td>
<td>0</td>
<td>0</td>
<td>8 (0.4%)</td>
<td>0</td>
</tr>
<tr>
<td>MSS/pMMR</td>
<td>392 (62%)</td>
<td>1836 (75%)</td>
<td>471 (71.3%)</td>
<td>1529 (69.8%)</td>
<td>760 (85.5%)</td>
</tr>
<tr>
<td>MSI/dMMR</td>
<td>65 (10.3%)</td>
<td>210 (8.6%)</td>
<td>136 (20.6%)</td>
<td>246 (11.2%)</td>
<td>129 (14.5%)</td>
</tr>
<tr>
<td>unknown MSI status</td>
<td>175 (27.7%)</td>
<td>402 (16.4%)</td>
<td>54 (8.1%)</td>
<td>415 (19%)</td>
<td>0</td>
</tr>
<tr>
<td>Stage 1</td>
<td>76 (12%)</td>
<td>485 (19.8%)</td>
<td>0</td>
<td>5 (0.2%)</td>
<td>169 (19%)</td>
</tr>
<tr>
<td>Stage 2</td>
<td>166 (26.3%)</td>
<td>801 (32.7%)</td>
<td>394 (59.6%)</td>
<td>53 (2.4%)</td>
<td>317 (35.7%)</td>
</tr>
<tr>
<td>Stage 3</td>
<td>140 (22.2%)</td>
<td>822 (33.6%)</td>
<td>267 (40.4%)</td>
<td>1653 (75.5%)</td>
<td>370 (41.6%)</td>
</tr>
<tr>
<td>Stage 4</td>
<td>63 (10%)</td>
<td>337 (13.8%)</td>
<td>0 (0%)</td>
<td>268 (12.2%)</td>
<td>33 (3.7%)</td>
</tr>
<tr>
<td>Stage unknown</td>
<td>187 (29.5%)</td>
<td>3 (0.1%)</td>
<td>0</td>
<td>211 (9.7%)</td>
<td>0</td>
</tr>
<tr>
<td>Left-sided CRC</td>
<td>248 (39.2%)</td>
<td>1607 (65.6%)</td>
<td>280 (42.3%)</td>
<td>1158 (52.9%)</td>
<td>487 (54.8%)</td>
</tr>
<tr>
<td>Right-sided CRC</td>
<td>176 (27.8%)</td>
<td>819 (33.5%)</td>
<td>375 (56.7%)</td>
<td>754 (34.4%)</td>
<td>332 (37.3%)</td>
</tr>
<tr>
<td>Unknown side</td>
<td>209 (33%)</td>
<td>22 (0.9%)</td>
<td>6 (1%)</td>
<td>278 (12.7%)</td>
<td>70 (7.9%)</td>
</tr>
</tbody>
</table>

Table S2: Clinico-pathological features of histopathological cohorts in our study. Abbreviations: colorectal cancer (CRC), number of patients (N Patients), interquartile range (IQR), microsatellite instability (MSI), microsatellite stability (MSS). Definitions: Right-sided CRC from cecum to transverse colon.