View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by Via Sapientiae: The Institutional Repository at DePaul University

DEPAUL UNIVERSITY
UNIVERSITY LIBRARIES

DePaul University
Via Sapientiae

College of Computing and Digital Media Dissertations

College of Computing and Digital Media

Spring 5-21-2021

Federated learning in gaze recognition (FLIGR)

Arun Gopal Govindaswamy
DePaul University, AGOVIND2@depaul.edu

Follow this and additional works at: https://via.library.depaul.edu/cdm_etd

Part of the Databases and Information Systems Commons, and the Data Science Commons

Recommended Citation

Govindaswamy, Arun Gopal, "Federated learning in gaze recognition (FLIGR)" (2021). *College of Computing and Digital Media Dissertations*. 35.
https://via.library.depaul.edu/cdm_etd/35

This Thesis is brought to you for free and open access by the College of Computing and Digital Media at Via Sapientiae. It has been accepted for inclusion in College of Computing and Digital Media Dissertations by an authorized administrator of Via Sapientiae. For more information, please contact digitalservices@depaul.edu.

FEDERATED LEARNING IN GAZE RECOGNITION

BY

ARUN GOPAL GOVINDASWAMY

A THESIS SUBMITTED TO THE SCHOOL OF COMPUTING, COLLEGE OF COMPUTING

AND DIGITAL MEDIA OF DEPAUL UNIVERSITY

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF SCIENCE IN DATA SCIENCE

DEPAUL UNIVERSITY

CHICAGO, ILLINOIS

2021

DePaul University
College of Computing and Digital Media

MS Thesis Verification

This thesis has been read and approved by the thesis committee below according to the requirements of the School of Computing graduate program and DePaul University.

Name: Arun Gopal Govindaswamy

Title of dissertation: Federated Learning in Gaze Recognition (FLIGR)

Date of Dissertation Defense: May 21, 2021

Advisor*

Dr. Jacob Furst

1st Reader

Dr. Daniela Stan Raicu

2nd Reader

Dr. Ilyas Ustun

3rd Reader

4th Reader (if applicable)

5th Reader (if applicable)

** A copy of this form has been signed, but may only be viewed after submission and approval of FERPA request letter.*

CONTENTS

ABSTRACT

The efficiency and generalizability of a deep learning model is based on the amount and diversity of training data. Although huge amounts of data are being collected, these data are not stored in centralized servers for further data processing. It is often infeasible to collect and share data in centralized servers due to various medical data regulations. This need for diversely distributed data and infeasible storage solutions calls for Federated Learning (FL). FL is a clever way of utilizing privately stored data in model building without the need for data sharing. The idea is to train several different models locally with same architecture, share the model weights between the collaborators, aggregate the model weights and use the resulting global weights in furthering model building. FL is an iterative algorithm which repeats the above steps over defined number of rounds. By doing so, we negate the need for centralized data sharing and avoid several regulations tied to it. In this work, federated learning is applied to gaze recognition, a task to identify where the doctor's gaze at. A global model is built by repeatedly aggregating local models built from 8 local institutional data using the FL algorithm for 4 federated rounds. The results show increase in the performance of the global model over federated rounds. The study also shows that the global model can be trained one more time locally at the end of FL on each institutional level to fine-tune the model to local data.

1. INTRODUCTION

Advancements in data-driven machine learning [1 ,2] have inspired various industries to collect volumes of data and perform predictive modelling. Healthcare is one such domain to adopt machine learning and deep learning [3-5]. Various innovations and advancements in artificial intelligence (AI) research have led to disruptive healthcare technologies [6, 7]. The huge volumes of data available for modelling inspires various machine learning solutions. These data are often collected, stored, processed, and modelled locally and in a decentralized manner [8, 9]. While these decentralized and local models serve the purpose for local institutions, these AI solutions are often applicable only to a subset of group and lack diversity. These locally available de-centralized data often follow different distribution and are subjected to biases (demographics like age, gender, geographic location or technical imbalances like acquisition, protocol, equipment manufacturing) affecting the performance of the AI models. These local data collected and used within single institutional sites such as clinics, laboratories, or research institutes restricts data diversity. Hence, collaboration and communication between institutions across national borders and research teams are increasingly becoming a necessity. Collaborative data sharing (CDS) techniques allows data collected by individual institutions could be shared, processed, and modelled. These shared data come from different distributions and hence, would add data diversity and attribute to better AI models.

While the idea of data sharing may sound straight forward, CDS has its own challenges. Using and sharing medical data, however, is restricted by regulatory systems that protect privacy. Depending on the context in which they are used and how they are related to other information, medical data, whether processed by healthcare providers in the delivery of healthcare or by researchers in the furtherance of generalizable knowledge, can be highly sensitive [10]. Studies have shown that most adults are concerned about the security and privacy of their [medical] data, and such concerns are associated with an increased likelihood of non-disclosure of sensitive information to a healthcare professional [11]. The vast number of data regulation and privacy policies limit data sharing capabilities. As of 2015, at least 109 countries in the world had data privacy laws in force [12]. In the United States, data storage, access, and sharing of medical and personal information of any individual is addressed in the HIPAA Privacy Rule. The HIPAA Security Rule outlines national security standards to protect health data created, received, maintained, or transmitted electronically [13]. The HITECH Act supports the enforcement of HIPAA requirements by introducing penalties for health organizations that violate HIPAA Privacy and Security Rules. Any institution that deals with protected health information must ensure that all the required physical, network, and process security measures are in place and followed.

These strict policies in place make it close to impossible to efficiently share data. Alternatively, federated learning is a machine learning setting where multiple entities (clients/ institutions) collaborate in solving a machine learning problem, under the coordination of a central server or service provider. Each client's raw data is stored locally and not exchanged or transferred; instead [8]. Federated Learning is seen as the future of digital health and solution for de-centralized collaborative learning [14 – 19]. This directly addresses the concerns with data sharing and privacy regulations. By local training of models and aggregation to produce global models, the methodology is simplified tremendously. Only the model weights are shared between them and research [30-34] shows the effectiveness of federated learning over traditional centralized training approaches.

In this work, a gaze recognition model is built on multi-institutional data using federated learning. In gaze recognition, Jocelyn [20] and Govindaswamy [21] argued for the need of convolutional neural networks. The learned CNN features outperformed traditional hand-crafted features in predicting physician gaze and is seen as the way to go forward. Fortunately, federated learning and deep learning goes hand-to-hand and works well. In this work, we train a single global model on the multi-institutional data without any centralized data sharing. A centralized federated learning topology is followed where model weights from the local institutions is aggregated centrally on a server. The contribution of this work is 2-fold: (1) Implementation and evaluation of the federated learning algorithm on gaze recognition; (2) additional local training after federated rounds to fine-tune local models; The research questions focused on this work are as follows: Does the global model built using global data using federated learning improve its performance over federated rounds? Do the local institutions gain by taking part in the federation and improve its performance over federated rounds? Can the local models benefit more by further fine-tuning of the global model using local data at the end of federated learning?

2. Background

2.1. Gaze Recognition

Gutstein et al. [22] - [24] used video recorded patient-physician interactions to extract motion information of the physician and the patient through optical flow algorithm [25] and You Only Look Once (YOLO) algorithm [26] to predict physician gaze. Gutstein et al., studied 6 interactions each from 2 doctors and 5 interactions from another doctor adding up to a total of 17 interactions. Three doctor- specific models were built using an AdaBoost algorithm and reported high performance in predicting physician gaze. The work posed several limitations due to the nature of clinical settings and camera angle. The most common issue was that of the doctor missing from one of the camera views which resulted in loss of up to 76% of frames from analysis and resulted in low generalizability to other videos capturing interactions of these three doctors with other patients. Another limitation of the work was the performance of these doctor-specific models on interactions from other doctors. Although the doctor specific models presented by Gutstein et al. produced high performing results, these models did not generalize well on clinical interactions which included a different doctor.

Govindaswamy et al. [27] extended the work by Gutstein [22 - 24] by including video interactions from higher number of doctors. The analysis was done on the image sequences to build an annotation tool for gaze recognition. Like reference [22 -24], Govindaswamy et al. [27] also used YOLO and optical flow algorithm in predicting physician gaze. The study worked on improving the generalizing capability of the gaze recognition model and built a model with an accuracy of 83%. The study also highlighted the efficiency of the model in labelling unseen new data for doctor gaze.

Govindaswamy et al. [21] studied different set of features and different classifiers in gaze recognition. The study used a traditional approach of image classification where features were manually extracted [22 – 24, 27] and the hand-crafted optical flow features were used to train a random forest [28] classifier. In another approach, the study extracted CNN-based learned features and used an end-to-end CNN model in gaze recognition. The results proved that the learned features are better than hand-crafted features in modelling diversely distributed data. The study also demonstrated the efficacy of the feature mapping in CNN by building hybrid architectures. The learned features were then used to build two classifiers – random forest [28] and k-nearest neighbor [29] algorithm. The two hybrid models were compared against an end-to-end CNN model and the results showed similar performance for the three models. The study concluded that the efficacy of the CNN comes from the feature learning for gaze recognition and that the choice of classifier has very less impact on the performance.

Tan et al. [20] aimed to support automated analysis of patient-physician interaction using deep convolutional neural network with transfer learning. Tan used 15 recorded video interaction between patients and doctors and built a deep CNN model by making use of transfer learning. Tan was able to build a robust gaze recognition model with an accuracy of 98% and showed the efficacy of the model in reducing the time and effort that goes into manual gaze rating. The study also showed the model's decision-making process by visualizing using the Grad Cam algorithm.

CNNs-based methods show better accuracy than conventional appearance methods in Gaze Estimation [45 – 55]. Zhang et al. [45] proposed a CNNs-based method to estimate gaze, the method was designed based on LeNet [46] and estimates gaze from eye images. Yu et al. [47] proposed a multitask gaze estimation model with landmark constrain, they estimate gaze from eye images. Fischer et al. [48] extracted feature from two-eye images with VGG-16 [49] to estimate gaze, they use an ensemble scheme to increase robustness of proposed method. Cheng et al. [50] proposed a CNNs-based network which uses two-eye images as inputs and utilizes the two-eye asymmetry to optimize whole network. Meanwhile, recent studies prove face images is effective in CNNs-based methods. Krafka et al. [51] implemented the CNNs-based gaze tracker in the mobile devices, it estimates gaze from face and eye images. Zhang et al. [52] proposed a spatial weights CNN to estimate gaze from face images. Deng et al. [53] proposed a CNNs-based method with geometry constraints, it uses face and eye images as inputs and can estimate gaze in free-head setting. Zhao et al. [54] proposed a CNNs-based method using dilated convolution to estimate gaze from face and eye images. Xiong et al. [55] combines the mixed effects model with CNN and estimates gaze from face images.

2.2. Federated Learning

Sheller et al. [30] experimented different data private collaborative learning approaches such as Federated Learning (FL), Institutional Incremental Learning (IIL), and Cyclic Institutional Incremental Learning (CIIL) on multi-institutional data. This study highlights the effectiveness of federated learning, a novel paradigm for data-private multi-institutional collaborations, where model-learning leverages all available data without sharing data between institutions, by distributing the model-training to the data-owners and aggregating their results. The study shows that federated learning among 10 institutions results in models reaching 99% of the model quality achieved with centralized data and evaluate generalizability on data from institutions outside the federation.

Rieke et al. [31] has presented federated learning as a future of digital health. Federated learning is shown as a promising approach to obtain powerful, accurate, safe, robust, and unbiased models. Various types of FL workflows were discussed briefly in the study and the promise of FL in the upcoming decade is highlighted. The study also discussed the algorithmic considerations in aggregating individual models using centralized topology. The study explains the impact of FL on stakeholders such as clinicians, patients, hospitals, researchers, AI developers, healthcare providers, and manufacturers. The study defines federated learning and discusses the challenges and considerations in adopting with federated learning approach.

Sheller et al. [32] used federated learning and two other collaborative learning methods such as Institutional Incremental Learning (IIL), and Cyclic Institutional Incremental Learning (CIIL) in building semantic segmentation models. The study shows ways of enabling deep learning models without sharing patient data. The results from the study shows that the federated model performs similar to that of models trained on shared data and quantitatively presented using the dice coefficient. The study also highlights the failure of IIL and CIIL models in multi-institutional collaborative learning. The study presents federated learning as benchmark approach to improve trained models without the need to share patient data, thereby overcoming potential privacy and data ownership concerns.

Li et al. [33] built a federated learning model in brain tumour segmentation using an aggregation algorithm. The focus of the paper was however in building a privacy-securing federated learning model. The authors pointed out the privacy loopholes in federated learning. Model inversion techniques could be employed to reconstruct training examples from the shared model weights, and this poses a concern in the federated learning research. The authors used the BraTS 2018 dataset to implement federated learning. The authors built a privacy-preserving federated algorithm by injecting noise to node's training process, distort the updates to limit the granularity of information shared. A centralized approach was also implemented and used as a baseline. The results show that the federated learning model produced comparable performance with a centralized approach and also implementing a privacy-preserving efficient algorithm.

Sinchhean et al. [34] implemented a collaborative learning approach (federated learning) in pneumonia detection. The task was to simply classify between normality and abnormality. The federation used 5 clients and 1 server in implementing FL. The evaluation was done on two different unseen data. A conventional data sharing model was trained separately and used as a baseline. The results showed an increase of 1.62% on the validation data when using the federated learning model.

3. Collaborative learning Techniques

3.1. Collaborative Data Sharing

In collaborative data sharing (CDS), the data from multiple institutions are sent to a server and a model is built and trained on a single centralized server. As shown in Figure 1, data from institution A and B are sent to a centralized server Institution C. A deep learning model is initialized, built, and trained on Institution C server using the shared data from Institution A and B. In this approach, only one machine is used to train the data. The trained model is then shared between institutions for usage. The idea may sound very simple; however, centralizing data involves regulatory, ethical, legal, and technical challenges [9 - 11]. CDS between many collaborators is challenging because of the privacy and data-ownership concerns [35, 36] especially, international collaborations.

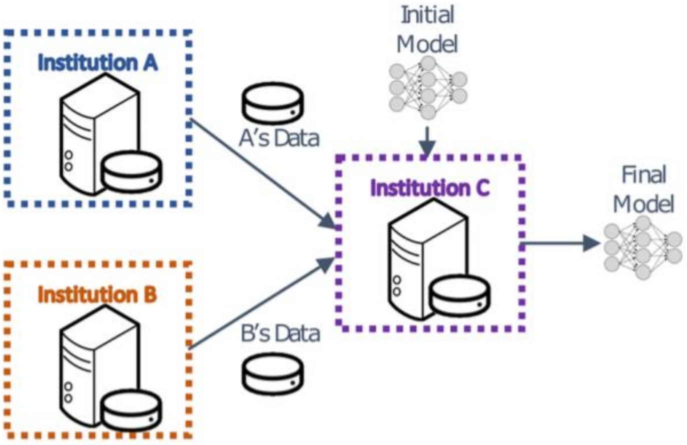

FIGURE 1: ARCHITECTURE OF COLLABORATIVE DATA SHARING

3.2. Institutional Incremental Learning

Institutional incremental learning (IIL) is a technique where each institution performs model training sequentially. The institutions undergo training phase sequentially as shown in Figure 2. A model is initialized and first trained on a single institution data. This trained model is then passed on to the next institution and is again trained. Likewise, the model is passed on to subsequent institutions and trained. The final model will be a result of a continuously trained model on individual institutional data. The difference between CDS and IIL is that the data is shared in the former approach while the model is shared in the latter. It is very clear that the IIL is superior in data handling, as this approach negates the challenges involved in data sharing. However, the performance is affected in IIL approach. The model performance drops in the case of IIL, when compared to CDS. Also, the final model created is biased towards last training institution data.

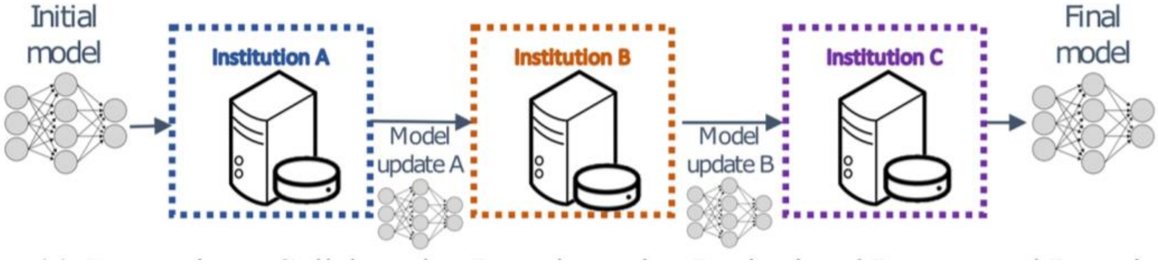

FIGURE 2: ARCHITECTURE OF INSTITUTIONAL INCREMENTAL LEARNING

3.3. Cyclic Institutional Incremental Learning

Cyclic institutional incremental learning (CIIL) shown in Figure 3 is very similar to Institutional incremental learning (IIL), whereas CIIL repeats IIL to a certain number of cycles. Like IIL, CIIL also deals with the problem of “catastrophic forgetting” [37], where the trained model favors the data it had most recently seen. The repeated cycles of training and limited epochs per institution helps the model gradually improve, reducing catastrophic forgetting, and results in better models than IIL. However, CIIL produced model still struggles with the bias.

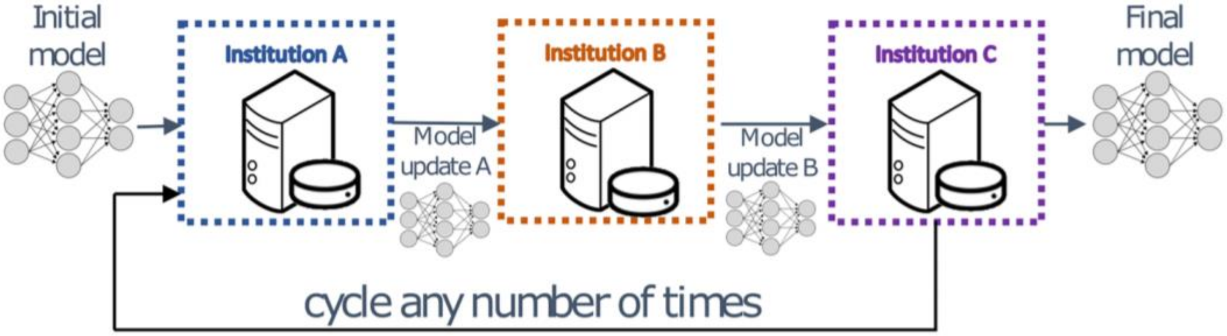

FIGURE 3: : ARCHITECTURE OF CYCLIC INSTITUTIONAL INCREMENTAL LEARNING

3.4. Federated Learning

Like IIL and CIIL, the model is shared between institutions in Federated learning (FL) shown in Figure 4. Each institution trains a deep learning model in parallel using the local institutional data and shares the model weights with other institutions in the collaboration. The model weights are then aggregated and updated across all institutions. This local institutional training, aggregation and update constitutes to 1 federated learning round. The models are trained to a fixed number of federated rounds until the model performs well over multiple institutional data and outside institutions.

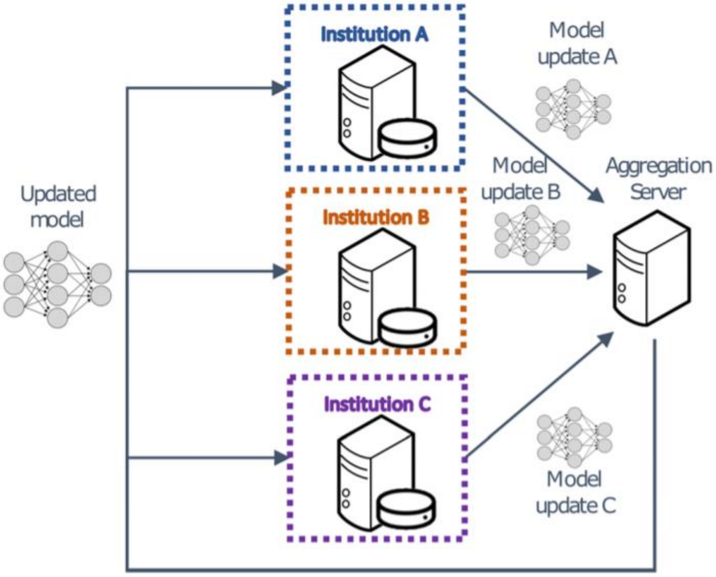

FIGURE 4: ARCHITECTURE OF FEDERATED LEARNING

Federated learning can be implemented through different topologies. Centralized, decentralized, hierarchical, hybrid hierarchical, sequential, aggregation [38 -40] and peer-to-peer [41, 42] are different topologies in federated learning. The methodology is shown in Figure 5. In this work, we would be implementing federated learning through a centralized aggregation server.

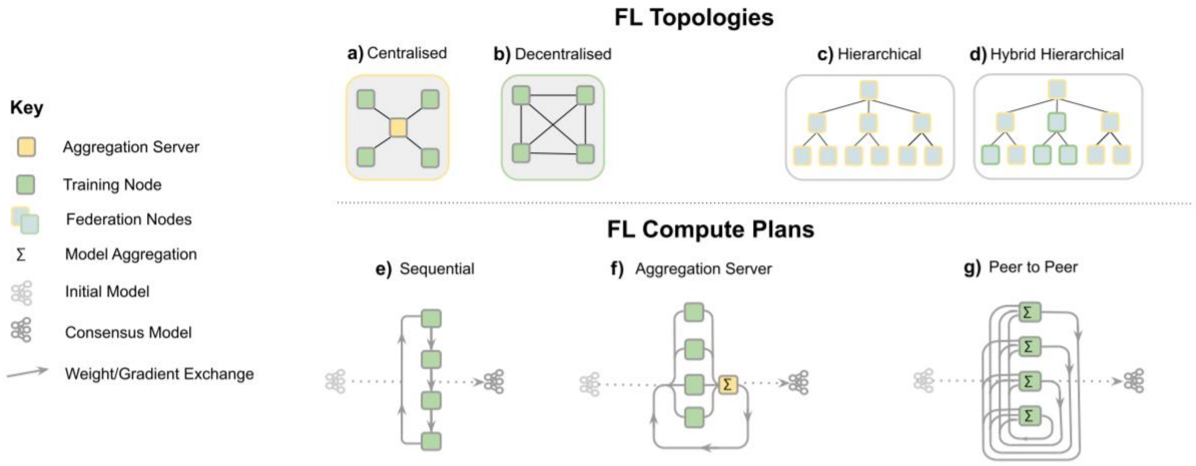

FIGURE 5: DIFFERENT TOPOLOGIES IN FEDERATED LEARNING

4. ARCHITECTURE

4.1. DATA

Current data consists of 101 recorded video interactions between patient and physician. The data was collected as a part of the study [43] consisting of 10 doctors and 101 patients. These videos were recorded in 5 primary care clinics through University of Wisconsin – Madison. Although these videos were a part of the same study and all the recorded videos follow same privacy policies and owned by single research group, we would consider these video recordings from each doctor as a separate institution. Video data from each doctor is considered as separate institution for the following reason – through previous study [21-24, 27] it was found that the features attributing to the video interactions between the doctors comes from different distributions. Hence, these video data pertaining to each doctor can be considered as a different institution. Also, this institutional data could help us in testing different collaborative learning methods, particularly federated learning. Table I shows the indices of the doctors and the interactions as per the original study.

TABLE I: SHOWN ARE THE INTERACTION INDICES DISTRIBUTION BY DOCTOR

Doctor Index	Interaction index distribution per physician
1	01, 02, 59, 66, 67, 73, 74, 89 - 91
2	03 - 08, 35 – 37
3	17, 21, 26 - 31, 39, 40
4	09 - 20, 25
5	22 - 24, 41 - 43, 51 – 54
6	32 - 34, 45, 48 - 50, 69, 80, 81
7	38, 44, 46, 47, 55 - 58, 61, 62
8	60, 63 - 65, 68, 71, 72, 75, 76, 92
9	70, 85 - 88, 93 – 97
10	77 - 79, 82 - 84, 98 - 101

4.2. Pre-processing

The interactions between doctors and the patients recorded as per the study [43] follows a video format. The study used 3 different cameras to capture the interaction – patient-centered camera, doctor-centered camera, and wide-angle camera. The patient-centered camera focuses on the patient, the doctor-centered camera focuses exclusively on the doctor, and the wide-angle camera gives a bird's-eye view of the entire room capturing both the doctor and the patient. All the 3 camera videos were then processed to a multi-channel view consisting of all the three cameras. Different views are showed in Figure 6.

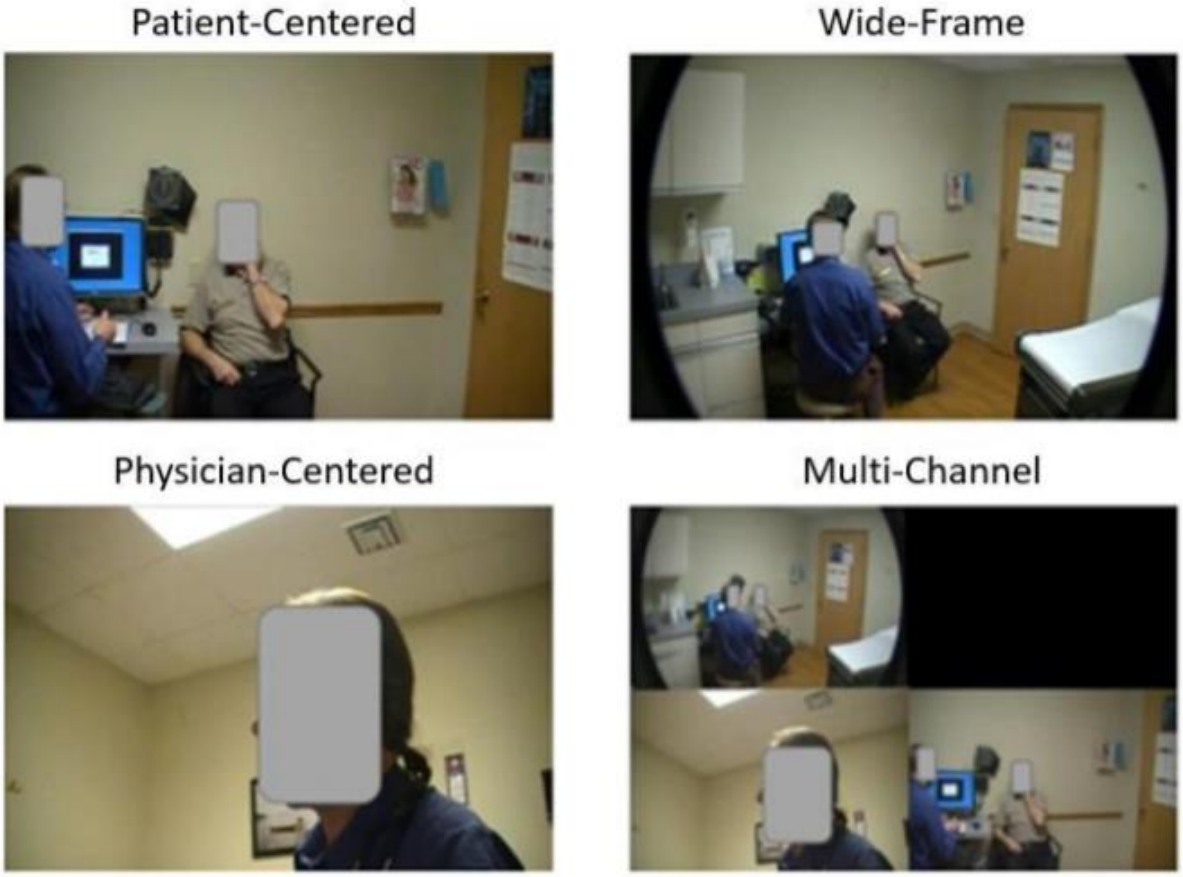

FIGURE 6: DIFFERENT CAMERA SETUPS CAPTURING THE DOCTOR AND THE PATIENT PRESENT IN A CLINICAL SETTING

4.2.1. Image Extraction

The videos were recorded at a frames per second of 29.97. It was found through previous study [20, 21] that doctor-centered videos are sufficient in predicting physician gaze. Hence, all the data-private collaborative models will use doctor-centered video data in training the models. However, human annotations of physician's gaze were provided pertaining to the multi-channel view. The annotation time were relative to the multi-channel video. Therefore, the multi-channel video was subjected to further image processing. The doctor-centered frame from the multi-channel video were cropped out to make sure the frames and the class labels are synchronized correctly.

4.2.2. Label Extraction

The labels originally provided by humans were categorized in sequences. For example, the video sequence from 00:01 seconds to 00:02 seconds were labelled "Class 0". We needed to convert the sequence-level labels to frame level. By applying the frames per second of 29.97, we converted the labels to frame-level. In this example, the 1:00 second was converted into 30th frame. Hence, frames from 30 to 60 were labelled as "Class 0".

4.2.3. Institutional data

Only 6 minutes of data from each interaction were used in model building. Hence, 6 minutes of data was chosen from each of the 101 interactions through manual analysis. While selecting the 6 minutes of data, it was made sure that the cropped data consists of a balanced labelling of data. While doing manual analysis, various anomalies were found in the data such as missing doctor from the clinic room, wrong-camera setup, etc., As a result, the number of interactions per institution were reduced as shown in the Table II.

TABLE II: SHOWN ARE THE INTERACTION INDICES PERTAINING TO EACH INSTITUTION

Institution	Interactions pertaining to each institution
1	3, 4, 6, 7, 35, 36
2	17, 26, 27, 28, 29, 31
3	9, 12, 13, 14, 15, 16
4	22, 23, 42, 43, 51, 52
5	32, 33, 34, 48, 49, 50
6	38, 46, 47, 55, 56, 57
7	60, 63, 64, 65, 68, 72
8	77, 78, 79, 84, 98, 100
9	1, 2, 59, 66, 67, 73, 74

101 interactions were reduced as shown in table. Each institution consists of 7 interactions each to maintain consistency across institutions. Institutions 1 through 8 were considered to be a part of the "Institutional Collaboration". Institution 9 was kept outside of the federated and was used for evaluation.

4.3. Federated Learning in Gaze Recognition

8 institutions were participating in federated learning and each institution consisted of 7 interactions. Of the 7 interactions, 1 interaction per institution was held out for internal validation. One held out interaction from each institution constitutes the validation dataset and was used for evaluating the global model. The validation data was further classified into internal and external validation data. The held-out interactions from each of the institution within federation was called as the internal validation and the interactions from outside of the federation was called as the external validation data. Table III shows the data used for training, testing, internal validation, and external validation from each institution.

TABLE III: SHOWN ARE THE INTERACTION INDICES PERTAINING TO TRAINING AND VALIDATION DATA

Institution	Interactions used for training and testing	Interactions used for internal validation	Interactions used for external validation
1	3, 4, 6, 7, 35, 36	37	-
2	17, 26, 27, 28, 29, 31	40	-
3	9, 12, 13, 14, 15, 16	19	-
4	22, 23, 42, 43, 51, 52	54	-
5	32, 33, 34, 48, 49, 50	80	-
6	38, 46, 47, 55, 56, 57	58	-
7	60, 63, 64, 65, 68, 72	75	-
8	77, 78, 79, 84, 98, 100	101	-
9	-	-	1, 2, 59, 66, 67, 73, 74

The global model trained and aggregated is evaluated against the validation data. 6 interactions from each institution constitute to model training and testing (Figure 7). Of the 6-minute from each interaction, the first 4 minutes of data is used for training and the next 2 minutes are used for testing. In federated learning, each of the 8 institutional model follow the same architecture. The models use Adam optimizer [44] and the batch size was 32. Figure 7 shows the interaction data split for training and validation.

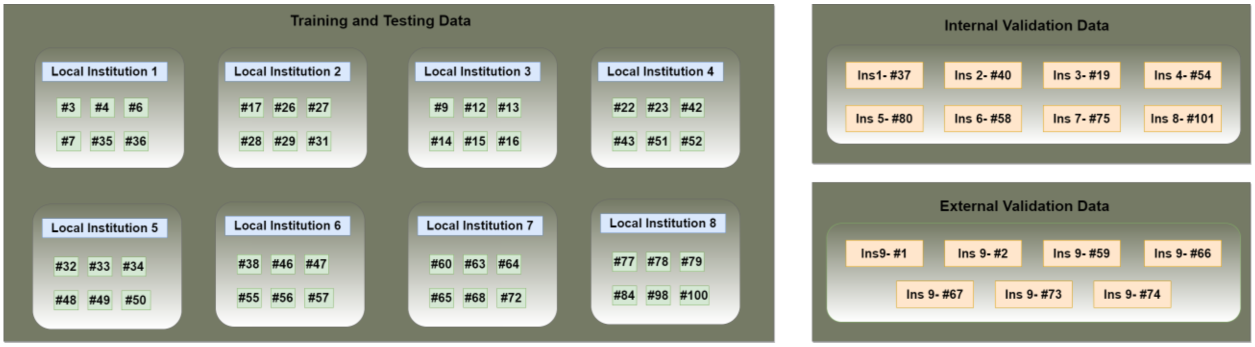

FIGURE 7: DATA SPLIT INTO TRAINING, TESTING, INTERNAL VALIDATION AND EXTERNAL VALIDATION SET

In Federated Learning algorithm, a global model is created by iteratively training local models and aggregating on the model weights. In this work, the centralized topology is implemented.

Algorithm 1. Example of a FL algorithm via Hub & Spoke (Centralized topology) with FedAvg aggregation.

Require: num_federated_rounds T

```
procedure AGGREGATING
    Initialize global model: $W^{(0)}$
    for t ← 1 ⋯ T do
        for client k ← 1 ⋯ K do ⊳ Run in parallel
        Send $W^{(t-1)}$ to client k
        Receive model updates and number of local training iterations from client's local training $\left(\Delta w_k^{(t-1)}, N_k\right)$ from client's local training.
        end for
    $W^{(t)} \leftarrow W^{(t-1)} + \frac{1}{\Sigma_k}\left(N_k \cdot \Delta W_k^{(t-1)}\right)$
    end for
    return $W^{(t)}$
end procedure
```

where w_k *denotes the weights of the local model*, Δw_k *denotes the change in weights*, *k=1, 2, … 8 denotes the index of the institutions, W denotes the global weights, t denotes the round of federation.*

As per the above algorithm, 8 local client models were trained on local institutional data. All the 8 local models followed the same network architecture as shown in Figure 8. The model follows VGG-16 and the weights were transformed using transfer learning. The bottom layers consisting of convolutional layers and max pooling layers were borrowed from VGG-16. The top layers were modified and the last convolutional layer from the VGG-16 was fine-tuned for better feature mapping. All the other layer weights were maintained. As a round zero of federated learning, all the 8 individual models were initialized with the same architecture as shown in figure, same hyper-parameters, and random initialized weights. The class labels fed into local models were also held the same. Having the same architecture and same class labels were the two constraints we held.

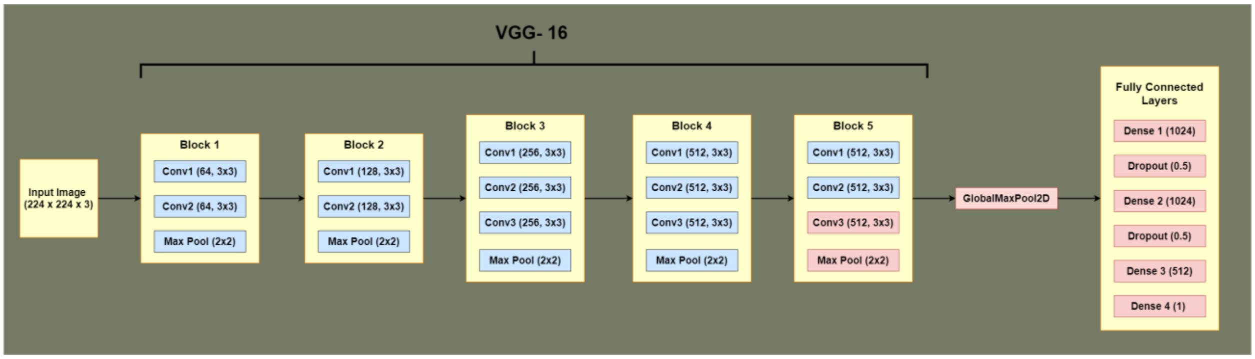

FIGURE 8: CONVOLUTIONAL NEURAL NETWORK ARCHITECTURE OF LOCAL MODELS

After complete training of the models using local institutional data, the model weights were then shared with the centralized server for aggregation. The centralized server then aggregated the model weights using the federated algorithm. The aggregated weights were then shared back to the local institutions for the next round of training. The local models were trained 4 times, creating a global model at the end of each federation round. These 4 global models were then evaluated against the validation data. The federated learning approach was held for 4 times due to time constraints. However, in an ideal scenario, the training would continue until a target is achieved or when the training becomes saturated and there is no further gain. A collaborative data sharing model, meaning a single model trained using all available data (global data) can also be used as a baseline and this performance can be set as a benchmark.

At the end of the federated learning rounds, the global weights were transferred to the local models, and the local models were trained one more time locally on local data. The hypothesis is that the local training would fine-tune the global model to local data and would boost the performance of the local institutions. These locally fine-tuned models were then evaluated against its local validation data and the external validation data as well.

4.4. Collaborative Data Sharing Model

Collaborative Data Sharing Model (CDS) is one in which data from all the institutions were gathered together and a single model was built. Data from Table III were again used for training, testing and validation of the model. In this scenario of CDS, a single model was built using the data from 8 different institutions. The model followed the same architecture shown in Figure 8. The model was run for 100 epochs with a batch size of 32 and using the Adam optimizer. The model was evaluated against internal and external validation data.

5. Results and Analysis

To evaluate the performance of our methodology, we used two different approach. In one approach, the global models were evaluated against each interaction in the validation data. Table IV shows the accuracy of the global models.

TABLE IV: PERFORMANCE OF THE GLOBAL MODELS ON INDIVIDUAL VALIDATION INTERACTIONS AT THE END OF EACH FEDERATED ROUND

	Performance of the global models at the end of federated rounds							
	Internal Validation							
	Ins #1	Ins #2	Ins #3	Ins #4	Ins #5	Ins #6	Ins #7	Ins #8
Global Model #1	79.48%	64.19%	81.67%	74.50%	61.74%	68.90%	67.14%	68.30%
Global Model #2	77.17%	69.98%	83.48%	74.23%	64.59%	73.87%	72.66%	73.28%
Global Model #3	82.98%	73.37%	79.70%	75.94%	63.39%	74.23%	77.49%	75.93%
Global Model #4	85.34%	75.20%	85.38%	76.90%	62.26%	74.99%	78.45%	80.45%

The result, overall, shows increase in the performance of the global models over federated rounds. This increase in performance also exhibits the effectiveness of the federated learning algorithm. The global models showed both increase and decrease in performance between each round across different validation data. However, overall, the global models improved and at the end of 4 rounds of federation, the performance was better than where than the starting performance. As a second approach, we evaluated the global models at the end of each round on the whole validation data. Table V shows the performance of the global models on the validation data.

TABLE V: PERFORMANCE OF THE GLOBAL MODELS ON INTERNAL VALIDATION DATA

	Global Model #1	Global Model #2	Global Model #3	Global Model #4
Accuracy	68.28%	72.49%	74.88%	75.25%
AUC Score	0.62	0.68	0.69	0.69

The results again showed that the performance increased through federated rounds, with small improvements. Both accuracy and AUC score show that the global models improved over federated rounds. After the end of 4 federated rounds, the global models were trained one more time locally on local data. This was called the fine-tuning of the model to local data. The performance of the local models without any federation, the global models and the fine-tuned models were studied. Figure 9 shows the performance comparison between different models. The blue dots were the local model trained and evaluated against its own validation data, orange dots were the global models evaluated against individual data, and the grey dots were the fine tune global models evaluated against its own validation data.

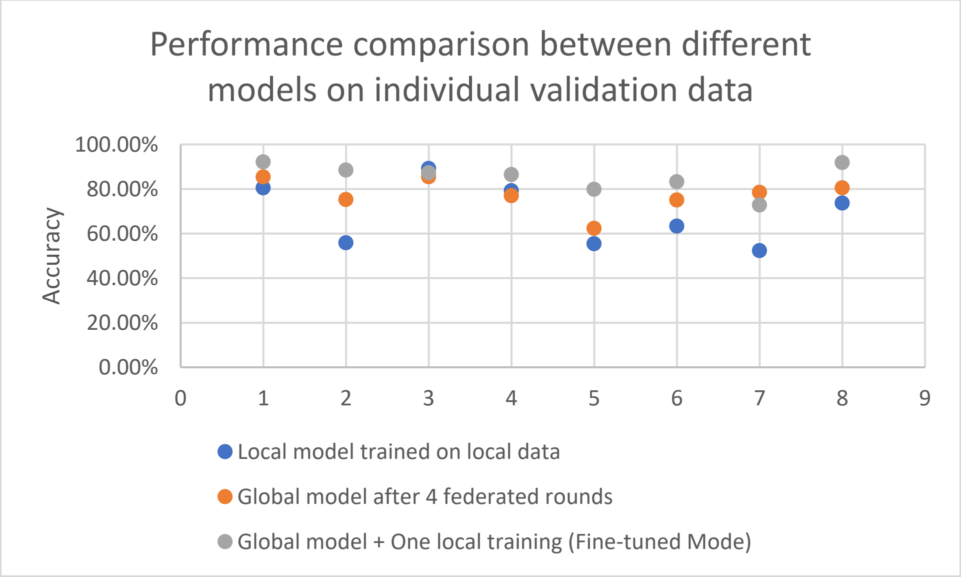

FIGURE 9: COMPARISON OF LOCAL, GLOBAL AND FINE-TUNED MODELS ON INDIVIDUAL VALIDATION DATA

The graph shows an overall trend in the performance of each institution's validation data. It can be noted that 7 of the 8 grey dots lie over the blue dots. This trend means that the local models predicted physician gaze with better accuracy after federated learning. The fine-tuned local models after federation perform better than the local models without any federated learning. The exception was in the Institution #3 model. The performance of the local model was higher than the global model and fine-tuned model. This was exception as the performance dropped after federated learning. Although, there was a 2% drop in the performance the model still performed with high accuracy of 87%. Also, it has to be noted that the orange dots stayed lower than the grey dots in 7 out of the 8 cases. This showed that the fine-tuned model saw an increase in the performance with one round of local training. Overall, evaluating the models against the validation set in different ways, the results showed that the global models kept improving with each additional round of federated learning. Also, additional fine-tuning showed increase in the performance of the global models on local data.

The global model and the fine-tuned models were also evaluated against the external validation data. Table VI shows the performance of the global models on external validation data. The results are centered around 50% and these performances would be rated poor given the task at hand was a binary classification.

TABLE VI: SHOWS THE PERFORMANCE OF THE GLOBAL MODEL ON EXTERNAL VALIDATION DATA

	Global Model #1	**Global Model #2**	**Global Model #3**	**Global Model #4**
Accuracy	47.65%	54.92%	51.39%	55.77%

Even though the global models performed well and improved over federated rounds on the internal validation data, the global models performed poorly on the external validation data. These results show that the institutions taking part in the federation can improve the models over federated rounds, while the institutions outside of the federation don’t. The global models trained on local data from institutions within federation using FL see improved performance over time and these global models are seen to perform well all the local institutional level validation data. This solidifies the importance of federated learning furthermore. Looking at the results, it can be safely concluded that institutions taking part in federated learning can reap benefits over time and how easily could the local institutions improve performance. Further, the fine-tuned models were also evaluated against the external validation data. Table VII shows the performance of the fine-tuned models on external validation data. As in the case of global models, the fine-tuned models also perform poorly on the external validation data. This further solidifies our understanding of federated learning algorithm.

TABLE VII: SHOWS THE PERFORMANCE OF THE FINE-TUNED MODELS ON EXTERNAL VALIDATION DATA

Fine-tuned Models	Accuracy
FT_Model_#1	48.89%
FT_Model_#2	54.96%
FT_Model_#3	58.71%
FT_Model_#4	54.22%
FT_Model_#5	46.37%
FT_Model_#6	32.11%
FT_Model_#7	53.64%
FT_Model_#8	52.33%

Further, **TABLE VIII** shows the performance of the collaborative sharing model on internal and external validation data. The CDS model was able to achieve a performance of 65.32% on the internal validation data, whereas the global model at the end of 4 federated round achieved a performance of 75.25% on the internal validation data. This result shows that the models trained using federated learning is more efficient than the traditional data sharing approaches.

TABLE VIII: PERFORMANCE OF THE COLLABORATIVE DATA SHARING (CDS) MODEL ON VALIDATION DATA

	Internal Validation	External Validation
Accuracy	65.32%	58.30%
AUC Score	.6938	.5489

Overall, the result show that the global model makes steady improvements in the performances over federated rounds and that the local institutions taking part in the federated have a significant gain when compared to the local institutions outside of the federation. The gain of local institutions measured by its performance on the local data also show a steady increase over federated rounds. It is also seen that fine-tuning of the global model with local data helps local institutions improve its performance.

6. CONTRIBUTIONS

The contribution of this work is thus 2-fold. One – federated learning algorithm was put into test in the gaze recognition problem. Institutional data from different local institutions were used to train local models and the local models were aggregated using the federated learning aggregation algorithm for 4 rounds of federation. The result provides evidence that the global models trained using federated learning is efficient than the local models and the collaborative data sharing models.

2nd contribution is the introduction of fine-tuning within federated learning approach. To our knowledge, nobody has ever fine-tuned the global models created using FL. In this study, the global model trained at the end of federated learning was trained one more time locally in local institutions on local data. This additional local training is called fine-tuning and the results show that the local institutions can not only improve their performance using federated learning, but also improve the performance using fine-tuning.

7. CONCLUSION AND FUTURE WORK

Federated learning algorithm was used as a solution to collaborative learning. Data available across multiple institutions were leveraged to build a robust gaze recognition model capable of predicting physician gaze across diversely distributed data. Federated learning approach was adopted, and local models were trained, aggregated, and evaluated against local held-out validation data. The study extensively evaluated the efficacy of the algorithm and showed improved performances with just 4 rounds of federated learning. The models were trained on decentralized data where no data was transferred between the institutions. The increase in the performance of the global models clearly showed the efficacy of the federated learning algorithm. The global models trained using federated learning were shown to improve over federated rounds of training. The institutions taking part in the federation were seen to benefit from the federated learning approach. The study also introduced a new paradigm of fine-tuned models in federated learning approach. The study showed that fine-tuning can help the local institutions to improve the performance. Through this study, the research questions on focus were answered as follows: global models make steady improvement over federated rounds and perform better than the local models and the collaborative data sharing model (CDS) as well; local institutions taking part in the federated has a significant gain when compared to the local institutions outside of the federation; fine-tuning of the global model with local data helps local institutions improve its performance.

One of the limitations of the work was limiting the number of federated rounds to 4. However, this was due to the time constraints. Other limitation was the constraints we held for local models – the local models had to follow the same architecture and same class labels for the federated learning to be performed. this study thus concludes by showing the ability to build diverse model without sharing of data. The concern with data sharing was that data storage and sharing is tightly regulated with strict policies in place. Although federated learning eliminates the need for data sharing reducing the risk of data breach, training data could be reconstructed from the global model weights through model inversion. The data used in this study has identifiable features regarding the patient and hence the transfer of model weights between collaborators poses as a risk. However, research shows that this caveat is being solved by disrupting model inversion by adding noise to the model weights. This is still an active field of research, as federated learning itself is still raw. The future work of this study would be in dealing with the privacy concerns regarding federated learning.

REFERENCES

1. Mitchell, Tom M. "Machine learning." (1997)., Jordan, Michael I., and Tom M. Mitchell. "Machine learning: Trends, perspectives, and prospects." Science 349, no. 6245 (2015): 255-260.
2. Mohri, Mehryar, Afshin Rostamizadeh, and Ameet Talwalkar. Foundations of machine learning. MIT press, 2018.
3. Aziz Sheikh, Kathrin M. Cresswell, Adam Wright, David W. Bates, Key Advances in Clinical Informatics, Academic Press, 2017, Pages 279-291, ISBN 9780128095232, https://doi.org/10.1016/B978-0-12-809523-2.00019-4.
4. Yu, Kun-Hsing, Andrew L. Beam, and Isaac S. Kohane. "Artificial intelligence in healthcare." Nature biomedical engineering 2, no. 10 (2018): 719-731.
5. Panesar, Arjun. Machine learning and AI for healthcare. Coventry, UK: Apress, 2019.
6. R. Bhardwaj, A. R. Nambiar and D. Dutta, "A Study of Machine Learning in Healthcare," 2017 IEEE 41st Annual Computer Software and Applications Conference (COMPSAC), 2017, pp. 236-241, doi: 10.1109/COMPSAC.2017.164.
7. Tresp, Volker, J. Marc Overhage, Markus Bundschus, Shahrooz Rabizadeh, Peter A. Fasching, and Shipeng Yu. "Going digital: a survey on digitalization and large-scale data analytics in healthcare." Proceedings of the IEEE 104, no. 11 (2016): 2180-2206.
8. Langer, Steve G. "Challenges for data storage in medical imaging research." Journal of digital imaging 24, no. 2 (2011): 203-207.
9. Shortliffe E.H., Barnett G.O. (2001) Medical Data: Their Acquisition, Storage, and Use. In: Shortliffe E.H., Perreault L.E. (eds) Medical Informatics. Health Informatics. Springer, New York, NY. https://doi.org/10.1007/978-0-387-21721-5_2
10. Dove E.S., Phillips M. (2015) Privacy Law, Data Sharing Policies, and Medical Data: A Comparative Perspective. In: Gkoulalas-Divanis A., Loukides G. (eds) Medical Data Privacy Handbook. Springer, Cham. https://doi.org/10.1007/978-3-319-23633-9_24
11. Agaku, I.T., Adisa, A.O., Ayo-Yusuf, O.A., Connolly, G.N.: Concern about security and privacy, and perceived control over collection and use of health information are related to withholding of health information from healthcare providers. J. Am. Med. Inform. Assoc. 21, 374–378 (2014)
12. Greenleaf, G.: Global data privacy laws 2015: 109 countries, with european laws now a minority. Priv. Laws Bus. Int. Rep. 133, 18–28 (2015)
13. M. authors, “HIPAA compliant hosting,” OnLINE TECH, Tech. Rep., 2015.
14. Sheller, M.J., Edwards, B., Reina, G.A. et al. Federated learning in medicine: facilitating multi-institutional collaborations without sharing patient data. Sci Rep 10, 12598 (2020). https://doi.org/10.1038/s41598-020-69250-1
15. Rieke, N., Hancox, J., Li, W., Milletarì, F., Roth, H. R., Albarqouni, S., Bakas, S., Galtier, M. N., Landman, B. A., Maier-Hein, K., Ourselin, S., Sheller, M., Summers, R. M., Trask, A., Xu, D., Baust, M., & Cardoso, M. J. (2020). The future of digital health with federated learning. NPJ digital medicine, 3, 119. https://doi.org/10.1038/s41746-020-00323-1
16. Xu, J., Glicksberg, B.S., Su, C. et al. Federated Learning for Healthcare Informatics. J Healthc Inform Res 5, 1–19 (2021). https://doi.org/10.1007/s41666-020-00082-4.
17. Kairouz, Peter, H. Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Keith Bonawitz et al. "Advances and open problems in federated learning." arXiv preprint arXiv:1912.04977 (2019).
18. T. Li, A. K. Sahu, A. Talwalkar and V. Smith, "Federated Learning: Challenges, Methods, and Future Directions," in IEEE Signal Processing Magazine, vol. 37, no. 3, pp. 50-60, May 2020, doi: 10.1109/MSP.2020.2975749.
19. Qiang Yang, Yang Liu, Tianjian Chen, and Yongxin Tong. 2019. Federated Machine Learning: Concept and Applications. ACM Trans. Intell. Syst. Technol.10, 2, Article 12 (February 2019), 19 pages. DOI: https://doi.org/10.1145/3298981

20. Tan T., Montague E., Furst JD., and Raicu DS., "Robust Physician Gaze Prediction Using a Deep Learning Approach". The 20th IEEE International Conference on BioInformatics And BioEngineering (BIBE 2020), Virtual Conference, October 26-28, 2020
21. Govindaswamy AG., Montague E., Raicu DS., Furst JD., "CNN as a feature extractor in gaze recognition", In 2020 3rd Artificial Intelligence and Cloud Computing Conference (AICCC 2020), Kyoto, Japan, December 18–20.
22. Gutstein, D.: 'Information Extraction from Primary Care Visits to Support Patient-Provider Interactions', DePaul University, 2020.
23. Gutstein, D., Montague, E., Furst, J.D., and Raicu, D.S.: 'Hand-Eye Coordination: Automating the Annotation of Physician-Patient Interactions', 2019 IEEE 19th International Conference on Bioinformatics and Bioengineering (BIBE), 2019, pp. 657-662.
24. Gutstein, D., Montague, E., Furst, J.D., and Raicu, D.S.: 'Optical Flow, Positioning, and Eye Coordination: Automating the Annotation of Physician-Patient Interactions', 2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), 2019, pp. 943-947.
25. Lucas, B. D. and T. Kanade. "An Iterative Image Registration Technique with an Application to Stereo Vision." International Joint Conferences on Artificial Intelligence, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1981, 674–679
26. Redmon, J., Divvala, S., Girshick, R., and Farhadi, A.: 'You Only Look Once: Unified, Real-Time Object Detection', 'Book You Only Look Once: Unified, Real-Time Object Detection' (2016, edn.), pp. 779-788
27. Govindaswamy AG, Montague E, Raicu DS, Furst J. "Predicting physician gaze in clinical settings using optical flow and positioning". In 2020 35th International Conference on Image and Vision Computing New Zealand (IVCNZ), November 25, 2020.
28. Breiman, "Random Forests", Machine Learning, 45(1), 5-32, 2001.
29. N. S. Altman (1992) An Introduction to Kernel and Nearest-Neighbor Nonparametric Regression, The American Statistician, 46:3, 175-185, DOI: 10.1080/00031305.1992.10475879
30. Sheller, M.J., Edwards, B., Reina, G.A. et al. Federated learning in medicine: facilitating multi-institutional collaborations without sharing patient data. Sci Rep 10, 12598 (2020). https://doi.org/10.1038/s41598-020-69250-1
31. Rieke, N., Hancox, J., Li, W., Milletarì, F., Roth, H. R., Albarqouni, S., Bakas, S., Galtier, M. N., Landman, B. A., Maier-Hein, K., Ourselin, S., Sheller, M., Summers, R. M., Trask, A., Xu, D., Baust, M., & Cardoso, M. J. (2020). The future of digital health with federated learning. NPJ digital medicine, 3, 119. https://doi.org/10.1038/s41746-020-00323-1
32. Sheller, M. J., Reina, G. A., Edwards, B., Martin, J., & Bakas, S. (2019). Multi-Institutional Deep Learning Modeling Without Sharing Patient Data: A Feasibility Study on Brain Tumor Segmentation. Brainlesion : glioma, multiple sclerosis, stroke and traumatic brain injuries. BrainLes (Workshop), 11383, 92–104. https://doi.org/10.1007/978-3-030-11723-8_9
33. Li, W. et al. Privacy-preserving federated brain tumour segmentation. In International Workshop on Machine Learning in Medical Imaging, 133–141 (Springer,2019).
34. Optimization and Implementation of a Collaborative Learning Algorithm for an AI-Enabled Real-time Biomedical System, Sinchhean Phea, Zhishang Wang, Jiangkun Wang, Abderazek Ben Abdallah, SHS Web Conf. 102 04017 (2021). DOI: 10.1051/shsconf/202110204017.
35. Tresp, V. et al. Going digital: a survey on digitalization and large-scale data analytics in healthcare. Proc. IEEE 104, 2180–2206. https://doi.org/10.1109/JPROC.2016.2615052 (2016).
36. Chen, M. et al. Privacy protection and intrusion avoidance for cloudlet-based medical data sharing. IEEE Trans. Cloud Comput. https://doi.org/10.1109/TCC.2016.2617382 (2016).
37. French, Robert M. "Catastrophic forgetting in connectionist networks." Trends in cognitive sciences 3, no. 4 (1999): 128-135.

38. Li, W. et al. Privacy-preserving federated brain tumour segmentation. In International Workshop on Machine Learning in Medical Imaging, 133–141 (Springer,2019).
39. Sheller, M. J., Reina, G. A., Edwards, B., Martin, J. & Bakas, S. Multi-institutional deep learning modeling without sharing patient data: a feasibility study on brain tumor segmentation. In International MICCAI Brainlesion Workshop, 92–104 (Springer, 2018).
40. Li, X. et al. Multi-site fmri analysis using privacy-preserving federated learning and domain adaptation: abide results. arXiv preprint arXiv:2001.05647 (2020).
41. Roy, A. G., Siddiqui, S., Pölsterl, S., Navab, N. & Wachinger, C. Braintorrent: a peerto-peer environment for decentralized federated learning. arXiv preprint arXiv:1905.06731 (2019).
42. Chang, K. et al. Distributed deep learning networks among institutions for medical imaging. J. Am. Med. Inform. Assoc. 25, 945–954 (2018).
43. Haskard, K.B., Williams, S.L., DiMatteo, M.R., Heritage, J., and Rosen- thal, R.: 'The Provider's Voice: Patient Satisfaction and the Content- filtered Speech of Nurses and Physicians in Primary Medical Care', Journal of Nonverbal Behavior, 2008, 32, (1), pp. 1-20
44. D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization,"arXiv preprint arXiv: 1412.6980, 2014.
45. Zhang, X.; Sugano, Y.; Fritz, M.; and Bulling, A. 2015. Appearance-based gaze estimation in the wild. In CVPR, 4511–4520.
46. Lecun, Y.; Bottou, L.; Bengio, Y.; and Haffner, P. 1998. Gradient-based learning applied to document recognition. Proceedings of the IEEE 86(11):2278–2324
47. Yu, Y.; Liu, G.; and Odobez, J.-M. 2018. Deep multitask gaze estimation with a constrained landmark-gaze model. In ECCV workshop.
48. Fischer, T.; Chang, H.; and Demiris, Y. 2018. Rt-gene: Realtime eye gaze estimation in natural environments. In ECCV.
49. Karen, S., and Andrew, Z. 2014. Very deep convolutional networks for large-scale image recognition. In CVPR.
50. Cheng, Y.; Lu, F.; and Zhang, X. 2018. Appearance-based gaze estimation via evaluation-guided asymmetric regression. In ECCV.
51. Krafka, K.; Khosla, A.; Kellnhofer, P.; Kannan, H.; Bhandarkar, S.; Matusik, W.; and Torralba, A. 2016. Eye tracking for everyone. In CVPR, 2176–2184.
52. Zhang, X.; Sugano, Y.; Fritz, M.; and Bulling, A. 2017a. It's written all over your face: Full-face appearance-based gaze estimation. In CVPR Workshop).
53. Zhu, W., and Deng, H. 2017. Monocular free-head 3d gaze tracking with deep learning and geometry constraints. In ICCV.
54. Chen, Z., and Shi, B. E. 2018. Appearance-based gaze estimation using dilated-convolutions. In ACCV.
55. Xiong, Y., and Kim, H. J. 2019. Mixed effects neural networks (menets) with applications to gaze estimation. In CVPR.