IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 2 March 2023, accepted 18 March 2023, date of publication 21 March 2023, date of current version 24 March 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3260027

TOPICAL REVIEW

A Systematic Review on Federated Learning in Medical Image Analysis

MD FAHIMUZZMAN SOHAN[1] **AND ANAS BASALAMAH**[2]
[1]Department of Software Engineering, Daffodil International University, Dhaka 1207, Bangladesh
[2]Department of Computer Engineering, Umm Al-Qura University, Mecca 21955, Saudi Arabia
Corresponding author: Md Fahimuzzman Sohan (fahimsohan2@gmail.com)

ABSTRACT Federated Learning (FL) obtained a lot of attention to the academic and industrial stakeholders from the beginning of its invention. The eye-catching feature of FL is handling data in a decentralized manner which creates a privacy preserving environment in Artificial Intelligence (AI) applications. As we know medical data includes marginal private information of patients which demands excessive data protection from disclosure to unexpected destinations. In this paper, we performed a Systematic Literature Review (SLR) of published research articles on FL based medical image analysis. Firstly, we have collected articles from different databases followed by PRISMA guidelines, then synthesized data from the selected articles, and finally we provided a comprehensive overview on the topic. In order to do that we extracted core information associated with the implementation of FL in medical imaging from the articles. In our findings we briefly presented characteristics of federated data and models, performance achieved by the models and exclusively results comparison with traditional ML models. In addition, we discussed the open issues and challenges of implementing FL and mentioned our recommendations for future direction of this particular research field. We believe this SLR has successfully summarized the state-of-the-art FL methods for medical image analysis using deep learning.

INDEX TERMS Federated learning, machine learning, medical image analysis, data privacy, systematic literature review.

I. INTRODUCTION

Image processing and analysis both are different tasks and often dependent on each other in terms of classifying an image data. To describe the image processing history we have to look quite back in 1973, an image of a Swedish model Lena is the first one that was used for image processing. Since then image processing has been applied in dozens of research fields, medical imaging is one of them. An image is essentially composed of 2D signals (vertical and horizontal), also with a number of pixels [1]. Different types of images have their different pixel parameters, during analysis these parameters help to extract respective information from the image. On the other hand, the task of the analysis part is to understand the processed images through different techniques, i.e., Machine Learning (ML); this technique includes different ML oriented algorithms. At the beginning, classical ML algorithms (e.g., SVM, naive bayes, decision tree) were used broadly in image processing research. Later on, it turned into neural network based modeling after introducing deep learning and now it is an integral part of any image analysis task including medical imaging. Every year the usage of medical imaging increases worldwide for diagnostics. The image data mainly represents various radiological images such as, X-ray, computed tomography (CT), magnetic resonance imaging (MRI), ophthalmology images, and so on. Besides, other data from eye, skin, cell have significant contributions in clinical imaging to detect, diagnose and treat diseases [2]. It is becoming increasingly important now to have these medical images being taken by different devices need to be sent across from one system to another and therefore they need a computer network. However, a large collection of such images creates a dataset, they are located and processed in cloud servers under ML approach.

In the era of AI, collaborative learning, more specifically sharing data among different institutions, multiple sources

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

TABLE 1. Contribution of available SLR on FL-driven medical data analysis and out study.

Contributions\Reference	[5]	[6]	[7]	[8]	Our Study
Area of investigation	All medical data	Healthcare area	Biomedical data	Oncology and cancer research	**All medical image data**
Number of articles	80	24	13	63	**17**
Demographic data presented	yes	yes	yes	no	**yes**
FL process disscused	Yes	yes	no	yes	**yes**
Applications of FL	no	yes	no	no	**yes**
Characteristics of federated datasets	yes	yes	no	no	**yes**
Additional privacy methods disscused	yes	yes	yes	yes	**yes**
Performance discussion	no	no	no	no	**yes**
Performance comparison with traditional ML	no	no	no	no	**yes**
Highlights	Elaborately discussed the FL process and privacy concern; but no discussion over performance.	All of the materials discussed broadly without performance comparison.	Keen about presenting survey data but very short discussion.	A comparative discussion of the architectures and learning approach between FL and general methods on cancer data.	**A SLR, covers all of the discussed issues in the FL oriented medical image analysis.**

can be very efficient in terms of building robust AI models. Since models are trained in centralized individual locations in traditional ML, the collaboration between models is quite tough. Contrastingly, X-rays, CT, MRI all of these are personal data pertaining individual patients which need to protect from risk of this medical information being disclosed or revealed to any unauthorized third party. In addition, even though data sharing is possible, the data store, processing, and analysis are still difficult tasks in a centralized manner. For such that scenario, data encryption-decryption could be a potential solution to exchange information between participants; however the process could be complex, time consuming and not sustainable [3]. So, instead of bringing the data to the location where the model is trained, why not bring the model to the data (institutions and the hospitals) and train directly there in-house, it allows collaborative learning without centralizing the dataset itself, this is called FL. It was first introduced in 2016 [4] and gained a lot of attraction within last couple of years for the healthcare domain. It addresses the privacy and data protection concern, which is currently an important problem in developing medical AI. In FL, the participants can train models locally and estimate different parameters for respective models, then share the parameters to a centralized server for aggregating them. Therefore, the focus is not on which data is used or what algorithms can be trained, the concept is managing the data in a different way where data privacy is reserved.

A. OBJECTIVE AND CONTRIBUTION

Since medical images are sensitive data, it needs to be protected and preserves the rights of users' personal information. We already discussed FL is arrived to solve the data privacy issue in collaborative ML and within the short time the concept has applied in different fields including medical imaging. Already many articles have been published on FL oriented medical image analysis and they successfully applied this unique data management technique in their research articles. At this stage, it is time to look back, need to review and assess what has been done till now, what are the impacts of FL on medical imaging. Meanwhile, some SLR have been published on the topic, however, they were about overall healthcare applications not particularly for the medical image analysis context. A SLR has been presented in [5], they considered all of the articles which have used all forms of medical data to train their FL models. Similarly in [3], [9], and [6] the authors have included the whole healthcare area to survey and review the papers. Some review articles presented specific medical domains, for example, Naeem et al. [10] worked particularly on brain tumor diagnosis using MRI images. Since FL is comparatively a new concept, most of the review articles emphasized on the design and implementation. Secondly, they discussed the privacy or security opportunity, which is the fundamental characteristic of FL. Some of them [5], [7], and [10] were formulated on different research questions, a common question was regarding the state-of-the-art FL methods; besides, data properties, impact, gaps and future research have been investigated. Alongside, several survey articles have been published on FL for healthcare informatics. Xu et al. [11] surveyed the papers that focus FL in the biomedical area to provide a review. Their effort was to summarize the privacy, statistical and system challenges that exist in this specific domain. A well-known article in this field [12], where the authors discussed the prime factors related to FL in digital health with challenges and solutions.

This study is a SLR, we exclusively investigated the FL in medical image analysis and extensively touched every component in the considered articles, specially the performance analysis and comparison with usual ML, which is the main distinction of our study corresponding to the previously published review papers. Our study consisted of several research questions and by answering the questions we illustrated the current research lay-out in the field of medical image processing using FL. In addition, several observations were discussed according to the findings extracted from the literature. Table 1

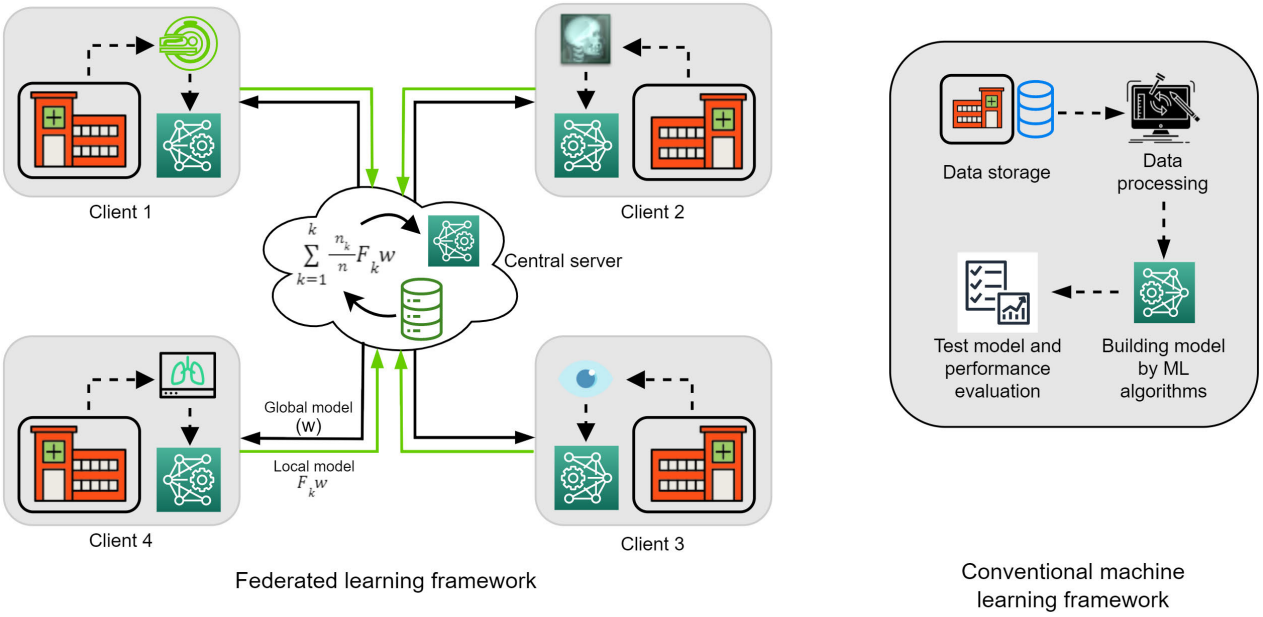


FIGURE 1. Two basic frameworks: working and communication flow of decentralised federated learning in left and usual machine learning in right for a hospital environment.

shows comparative analysis of our contribution and related review articles, our study explored the demographic data, FL architecture, privacy preserving concern, federated data management, and performance of FL models. We did not find any article which has worked particularly on medical images. Consequently, this study can be an outline for future research of FL application in medical imaging. The following are the key contributions of our paper:

- We surveyed the insights of FL solely in medical image research in a systematic way.
- We provided the latest implementation, advancement, and tendencies toward medical image analysis research using FL in different aspects.
- We presented and compared the performance of different FL architectures used in the reviewed articles with traditional ML models, which is the first of its kind.
- For incoming contributors we discussed open issues, challenges, and future direction of the research field.

Rest of the article is structured with six sections. Basic FL concept is introduced in Section II. Section III described the procedures of this review. The results of this investigation are presented through different research questions in Section IV. Open issues and challenges are discussed in Section V. Besides, Section VI includes the limitation of this study. Lastly, the conclusion and future directions is provided in Section VII.

II. FEDERATED LEARNING

In this section we have described an overview of FL architecture. The concept of FL is not related directly to the ML components, it is all about a data management process to share data between multiple clients in a privacy preserving manner. For a practical example, suppose a hospital environment that produces some data, also has a model and some computer resources that would like to tackle a specific problem by an AI system. Moreover, the dataset in the institution has not been sufficient to train the model which is able to address this problem. Another hospital dealing with similar difficulty wants to work together on this promise where they have a common goal and can solve a common task. However, both hospitals have different data locally and they need to use each other's data without sharing data directly. This collaborative model training without sharing the data is exactly the purpose of FL.

In Fig. 1, we have presented FL in left and traditional ML framework in right to illustrate the fundamentals of both for a hospital environment. In association with that, as supplementary information we have listed necessary keywords and their explanations related to decentralized FL implementation in Table 2. Since FL consists of multiple sources of data, we have shown four clients in the figure. Each of the clients has few common duties, they collect the data from the hospitals, train them using the local ML models and estimate some parameters. These parameters are sent to the central server from every client, not the data itself. Once the central server has received all the local modes' parameters, it aggregates them and takes the weighted average, this is known as the global model and sent back to all of the clients. By this process a learning round is completed and repeated for the next round.

However, a well known federated averaging algorithm is FedAvg [13], proposed by Google in 2016, it calculates

TABLE 2. Some commonly used components and their definitions in federated network.

Keyword	Explanation
Client	They are the participants of the collaboration process.
Local data	Each participant brings their own data which called local data.
Local model	Each participant trained a ML model using the local data.
Central server	An storage where learning results of all local models are collected.
Model aggregation	Aggregating all of the local modes.
Global model	This model is the average of the aggregated model.

weighted average of the individual clients. It is very expected that the data quantity could not be the same across the clients, sources with larger datasets will have correspondingly larger weighted losses, individual clients losses are minimized to an overall global loss which is called weighted average. Under FedAvg, every client trains a model for a defined number of epochs through Stochastic Gradient Descent (SGD) algorithm and transmits the learning parameters to the central server and the server performs aggregation in the form of an averaging. The mathematical presentation of FedAvg is:

$$f(w) = \sum_{k=1}^{k} \frac{n_k}{n} F_k w$$

In this function there are k number of clients and each client has its own loss function $F_k w$. Then weight each of the losses by the size of the client's dataset n_k. Hence, the overall objective is to minimize a global loss which is a weighted combination of local losses and the local loss is computed on private data which is never shared, only model updates are shared. Apart from the FedAvg, there are many research directions and varieties of FL going on such as SecAgg [54]. Though different combinations exist in the FL implementation, two characteristics are maintained expectedly: the datasets are distributed and remain local, not centralized and have a collaborative model to work towards the same goal.

III. RESEARCH METHOD

There are several review article types available in the literature to do deeper level of research, such as narrative reviews, systematic reviews. We mentioned at the beginning that a systematic review has been conducted for our investigation. Mainly two SLR methods are popular in practice, one is PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) and another is Kitchenham's guidelines; the second one is mainly considered in computer science and software engineering research fields [14]. To conduct this review we followed the PRISMA procedures which is the most common way of performing SLR in the healthcare sector [6]. However, for a SLR, first we need to identify

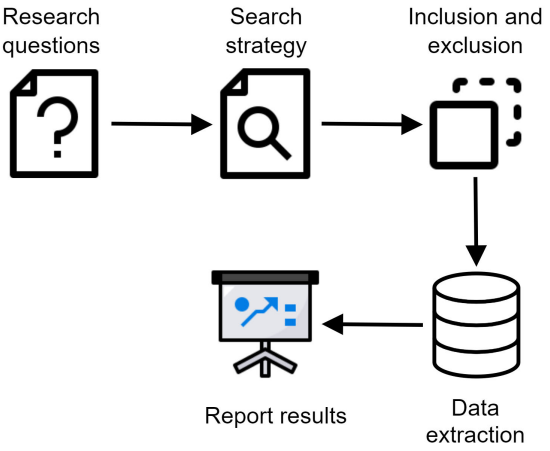


FIGURE 2. The steps taken to conduct this study.

relevant articles that focus on a very specific research area and question(s), secondly appraising the quality of the studies performed and the strength of the evidence in the papers, and lastly synthesize the findings to draw respective conclusions. Fig. 2 shows all of the steps taken to conduct this review sequentially.

TABLE 3. Formulated research questions of this review.

Context	Questions
Overview	**RQ1** What are possible applications of FL?
	RQ2 What problems were solved?
Dataset	**RQ3** What type of dataset used?
	RQ4 Are the number of data samples sufficient?
	RQ5 Are non-IID data distribution considered?
ML Framework	**RQ6** What type of ML algorithms are used?
FL Implementation	**RQ7** Are any additional privacy methods implemented?
	RQ8 What types federated data partitioning are used?
	RQ9 Which federated frameworks are used?
Experimental	**RQ10** What are the performance measures used in the studies?
	RQ11 How is the performance of the FL frameworks reported?
	RQ12 How perform the FL approach compared to the conventional models?

A. RESEARCH QUESTION

Our first step of this review was to establish a group of questions which will describe the literature in the most effective way. Table 3 shows the five contexts and their associated 12 research questions. First context is the overview that talked about the application and problem solved by the FL; next a broad explanation over the datasets was presented; third ML framework; then implementation of FL was discussed including privacy method, types of FL; and lastly the experimental substances, specially the performance comparison have been presented.

B. SEARCH PROCESS

Since FL was first presented in 2016, the search process of the review was limited over the time period from 1 January 2017 to 30 June 2022. We discovered all of the common databases considered by previous researchers; for example, Science Direct, IEEE Xplore digital library, Springer Link, Wiley Online Library, SPIE digital library, ACM digital library, Multidisciplinary Digital Publishing Institute (MDPI), Nature Portfolio, Taylor & Francis, and Google Scholar. The searching criteria is different across the platforms, we used advanced options of each database to search articles with Boolean "AND" and "OR" expressions. Our study focused on the implementation of FL in healthcare image processing, so that we carefully avoided the other applications. The search phrases looked over the titles, abstracts, and keywords in each of the databases. Fig. 3 depicts the PRISMA flow diagram where whole statistics of article consideration in this review has been presented. After the search operation, primarily collected articles have gone through a selection process, we have described them in the next sections.

C. INCLUSION AND EXCLUSION CRITERIA

Literature search strategy is a big challenge while it is needed to find too many papers, these circumstances are solved by a predefined inclusion and exclusion criteria in SLR. This might include limiting the search to only those that contain certain types of studies. However, the processes ensure the task achievement properly, reduce the possibility of bias and protect the selection process from irrelevant research documents. We implemented the inclusion and exclusion on the collected articles from the databases to reach the exact materials that are seeking the readers. We emphasized the following points to include articles for final analysis:

- Article that studied medical image datasets.
- ML model developed with the FL environment.
- FL was the main focus in the findings (result analysis/comparison).

Since we performed keyword search, the articles were collected based on the words present in the paper, even if it was mentioned for a single time. Therefore, we excluded the articles that are not relevant and does not fulfill our scope based on the given criteria:

- Articles that used private dataset(s) for the ML model.
- Studies that are not mainly focused on FL and medical image data.
- Hybridization or modify the theme of FL, e.g., federated reinforcement learning.
- Abstract, short article, any pre-print, any book or book part.
- Articles do not have a clear presentation of the results using ML based performance measures (e.g., [85], [86]).

The functionalities of inclusion and exclusion are observed in Fig. 3. It shows the number of initially collected articles from different databases is 161. We have removed the duplicate articles from there and 138 articles were taken for further steps. After that we screened the articles for two times under two different conditions, first we gently explored the title and abstract which helps to remove 96 articles, besides, we extensively investigated the full text of rest 42, where another 25 papers have been disqualified. Finally, we discovered 17 from 161 articles to hold our review.

D. DATA EXTRACTION

Data collection mostly involved in research questions of our study, we extracted information in order to cover the questions perfectly. At first we created a spreadsheet and input respective information headers on the top. We worked on the 17 articles individually, each time all of the information has been gathered distinctively on the spreadsheet and they were used as our findings. The following data are extracted from every articles:

1) Document title, publication year, and journal/ conference name.
2) Used datasets and their federated settings.
3) The security or privacy protocol used for FL.
4) The algorithms used to train ML models.
5) Performance of the FL model.

IV. RESULTS

We assembled this section following the research questions that we described in Section III. In the upcoming sections, first we have presented the demographic analysis (also known as numerical analysis) data along with the key contributions and limitations of each reference work in Table 4, thereafter we answered the 12 questions successively.

A. OVERVIEW

RQ1 What are possible applications of FL?

We found the application of FL in different research fields, such as, Diabetic Retinopathy (DR), MRI classification, cancer, pneumonia, COVID-19 detection, and few more. These topics are popular in medical image processing research with conventional ML. Hence, FL also creates new scope to research due to the privacy production efficiency which is essential for this particular imaging research.

In 2019, coronavirus disease hitted all over the world and created a crisis regarding identification of COVID-19 samples. The RT-PCR test is the most reliable diagnosis method of the diseases, since inadequate testing kits and some technical limitations, researchers tried to explore alternative ways of COVID screening. Therefore, hundreds of ML based automated and time saving COVID-19 detection models have been presented within the last two years [33]. ML based COVID analysis is mostly carried out by radiological chest images, i.e., X-ray and CT images. Among the contributions, FL also discussed and implemented several detection models as data privacy was a big concern there. In this study, we found six articles out of 17 were specifically worked on COVID-19 detection. Feki et al. [18] proposed a collaborative FL for COVID-19 screening from chest X-ray images; they cooperated with multiple medical institutions

TABLE 4. Numerical data of considered articles for this study which include publication year, name of publisher, and data analysis method.

Article	Year	Task	Contributions	Limitations
Lo et al. [16]	2021	*Both*	The article performed classification and segmentation both tasks using FL framework, also provided source code online.	Number of data samples is very small; non-IID data and clear discussion over experiment procedure are not considered.
Połap et al. [17]	2021	Classification	Blockchain based additional privacy used for federated models.	The accuracy is low; non-IID data is not considered
Feki et al. [18]	2021	Classification	The proposed FL framework is well described; non-IID data considered; results comparison is given.	Number of data samples is very small.
Połap and Woźniak [19]	2021	Classification	The proposed FL framework is well described; result comparison with state-of-the-art is given.	Non-IID data is not considered.
Li et al. [20]	2020	Classification	All segments of the experiment are well presented; provided source code online; additional privacy method is used.	Number of data samples is very small; non-IID data is not considered.
Yang et al. [21]	2021	Segmentation	A federated semi-supervised learning method is presented with sufficient experimental details.	Need to compare results with state-of-the-art methods.
Połap [22]	2021	Classification	A fuzzy consensus with the FL method proposed and all of the experiment procedures are presented properly.	Non-IID data is not considered.
Chakravarty et al. [23]	2021	Classification	FL applied to chest radiograph screening using deep learning methods.	Performance comparison with state-of-the-art methods is not given; non-IID data is not considered.
Zhang et al. [24]	2021	Classification	A dynamic fusion based vertical FL proposed for COVID-19 detection.	Non-IID data is not considered.
Qayyum et al. [25]	2021	Classification	A vertical FL based multi-modal COVID-19 diagnosis approach is presented with sufficient experimental details.	Non-IID data is not considered.
Linardos et al. [26]	2022	*Both*	Deep learning based FL model for cardiovascular disease detection; provided source code online.	Number of data samples is very small.
Kaissis et al. [27]	2021	Classification	All required sections and experimental details are presented; provided source code online; additional privacy method is used.	Non-IID data is not considered.
Adnan et al. [28]	2022	Classification	A collaborative ML based approach is proposed for tissue image data; all necessary sections are included.	Performance comparison with state-of-the-art methods is not given.
Yan et al. [29]	2021	Classification	FL for COVID-19 detection model using chest x-ray images.	Non-IID data is not considered.
Hashmani et al. [30]	2021	Classification	FL is used for skin disease detection using dermoscopy images.	Performance comparison with state-of-the-art methods is not given; Non-IID data is not considered.
Zhou et al. [31]	2022	Classification	Multi-institutional medical image classification performed with federated modeling; data heterogeneity considered.	Visualization and interpretation of learned model is not addressed clearly.
Salam et al. [32]	2022	Classification	Studied the efficacy of FL for COVID-19 detection using medical images.	Performance comparison with state-of-the-art methods is not given; Non-IID data is not considered; presentation style is not reader friendly.

Both - Classification and Segmentation

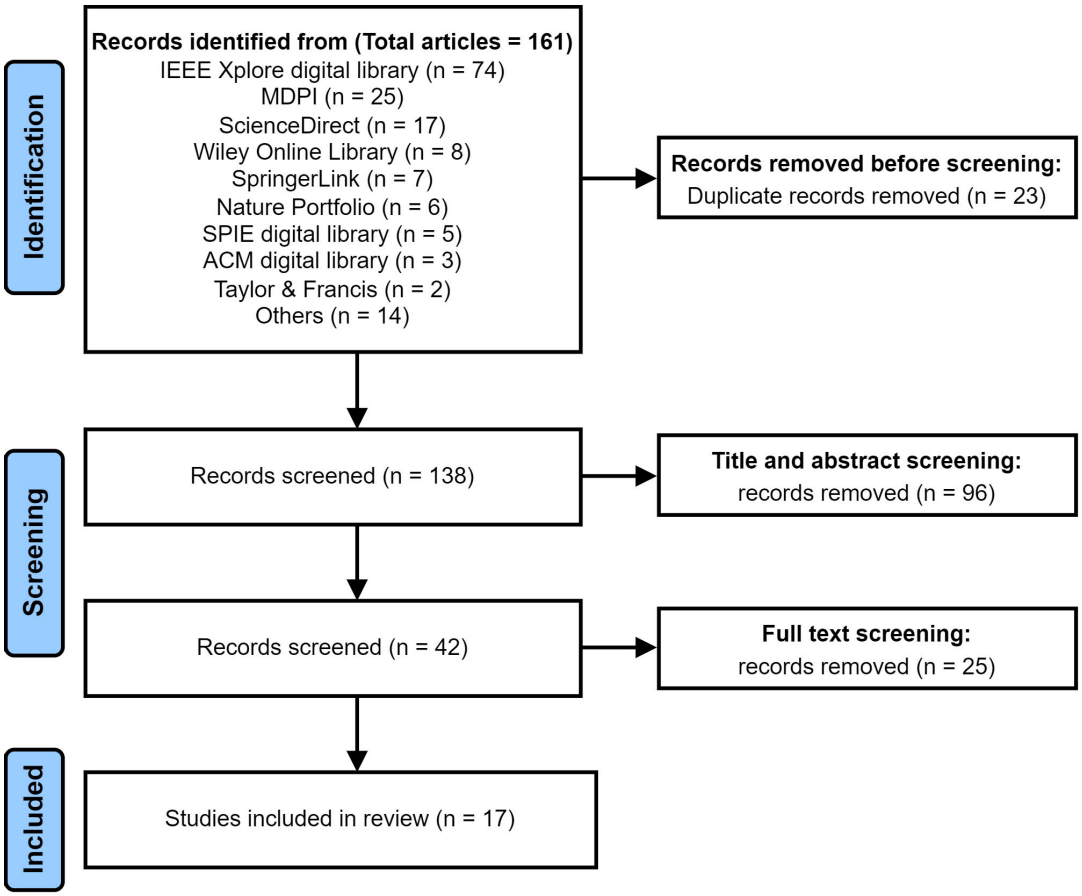


FIGURE 3. Article consideration process of this review according to PRISMA flow diagram.

without sharing their data. Similarly, Zhang et al. [24] and Yan et al. [29] used X-ray and CT image data for different Convolutional Neural Network (CNN) architectures in FL settings. References [21], [25], and [32] also have contributed to the COVID-19 infection in a multinational way. However, during the pandemic such that artificial intelligence tools were not clinically used significantly to diagnose COVID-19, all of them were experimental operations and hopefully the contribution will help in future initiative.

Millions of patients are suffering from fatal diseases worldwide, cancer is top of them. Researchers have shown early detection of cancer can save a large number of lives [34]. Consequently, deep learning has emerged as a potential of early cancer detection by the help of medical images. It extracts features from the raw images and provides decisions regarding cancer detection with notable performance. As a part of ML technique, FL has been considered in several cancer diagnosis techniques, Fig. 4 shows 29.4% articles (five out of 17) of this review were formed on cancer detection. Researcher Polap and their team have published three research papers [17], [19], [22], all of them focused on skin cancer detection with the FL environment. They used seven different skin marks (classes) to train the detection models and successfully implemented the privacy protected FL. Moreover, Hashmani et al. [30] applied FL on a series of dermoscopy images to classify nine different skin diseases. Nowadays, important internal organs of human body, such as lung, breast cancer are the leading causes of cancer death. A FL oriented lung cancer detection model has been proposed by Adnan et al. [28]. They demonstrated that their model achieved acceptable performance while decentralized data configuration applied.

One of the domains is Diabetic retinopathy (DR) analysis. Diabetes is a chronic disease that affects millions of people globally and uncontrolled diabetes can lead to serious damage to the body's system including eyes. DR is a common diabetic eye disease and the number one cause of vision loss and blindness in the world. It occurs when diabetes damages the small blood vessels on the retina. In the primary care clinic, those retinal images can be transmitted to an eye care specialist who investigates the image and then provides a consultation. However, these days deep learning algorithms can detect the DR within seconds with high accuracy. Lo et al. [16] analyzed the retinal images to classify the DR positive and non-DR samples using the FL approach. In another article, Zhou et al. [31] introduced a FL framework which classifies five scalability categories of DR, 0 to 4 (No DR to Proliferative DR).

TABLE 5. List of dataset used in federated medical imaging with references for quick access.

Article	Dataset Name	Total number of samples	Dataset References
[16]	1) SFU prototype swept-source, 2) RTVue XR Avanti, 3) Angioplex, and 4) PLEX Elite 9,000	153 retinal images	[35]
[17]	MNIST: HAM10000 dataset	10,015 dermoscopy images	[36]
[18]	1) Cohen JP dataset and 2) TB x-ray	216 lung X-ray images	[37], [38]
[19]	MNIST: HAM10000 dataset	10,015 dermoscopy images	[36]
[20]	ABIDE dataset	370 brain MRI images	[39]
[21]	*Private data*	2,317 chest CT images	*ND*
[22]	MNIST: HAM10000 dataset	10,015 dermoscopy images	[36]
[23]	CheXpert dataset	223,414 chest X-ray images	[40]
[24]	1) Qatar-Dhaka COVID-19 Data, 2) COVID-CT, and 3) Figure 1 dataset	746 CT and 2,960 X-ray images	[41]–[43]
[25]	1) Cohen JP dataset, and 2) Chest ultrasound	1564 X-ray and 545 ultrasound images	[88]
[26]	1) M&M and 2) ACDC datasets	180 heart MRI images	[44], [45]
[27]	1) Mendeley data, 2) MedNIST dataset, 3) MSD Liver Segmentation dataset	6,284 lung X-ray images	[46]–[48]
[28]	TCGA cancer data	2,580 lung tissue images	[49]
[29]	COVIDx	15,282 lung X-ray images	[50]
[30]	ISIC challenge dataset 2019	25,331 dermoscopy images	[51]
[31]	DR dataset	3,662 retinal images	[52]
[32]	1) Chest X-ray (CXR), and 2) COVID 2019 dataset	5,144 chest X-ray images	[41]–[43]

ND - No specific data was found

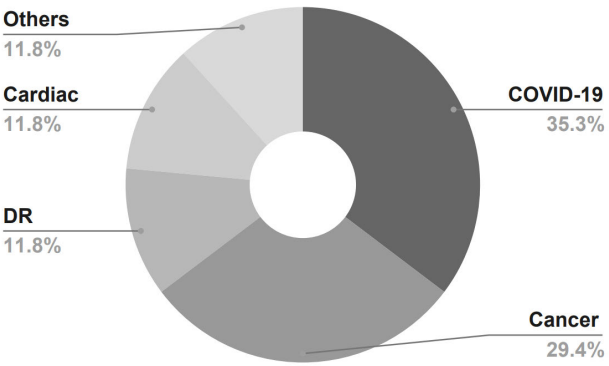


FIGURE 4. Percentage of FL applied in different diseases diagnosis research.

Linardos et al. [26] considered FL for Diagnosing Hypertrophic Cardiomyopathy (HCM), whether the subjects are suffering from HMC or normal. In addition to that, a multi-label cardiac diseases classification has been proposed by Chakravarty et al. [23], where 14 classes were examined. The other application includes Autism Spectrum Disorders (ASD) detection. Li et al. [20] applied deep learning in a FL environment to classify MRI images. Their model worked for identifying the ASD using the MRI analysis technique. We also found FL is used in pneumonia detection, Kaissis et al. [27] proposed a model that able to detect different pneumonia samples.

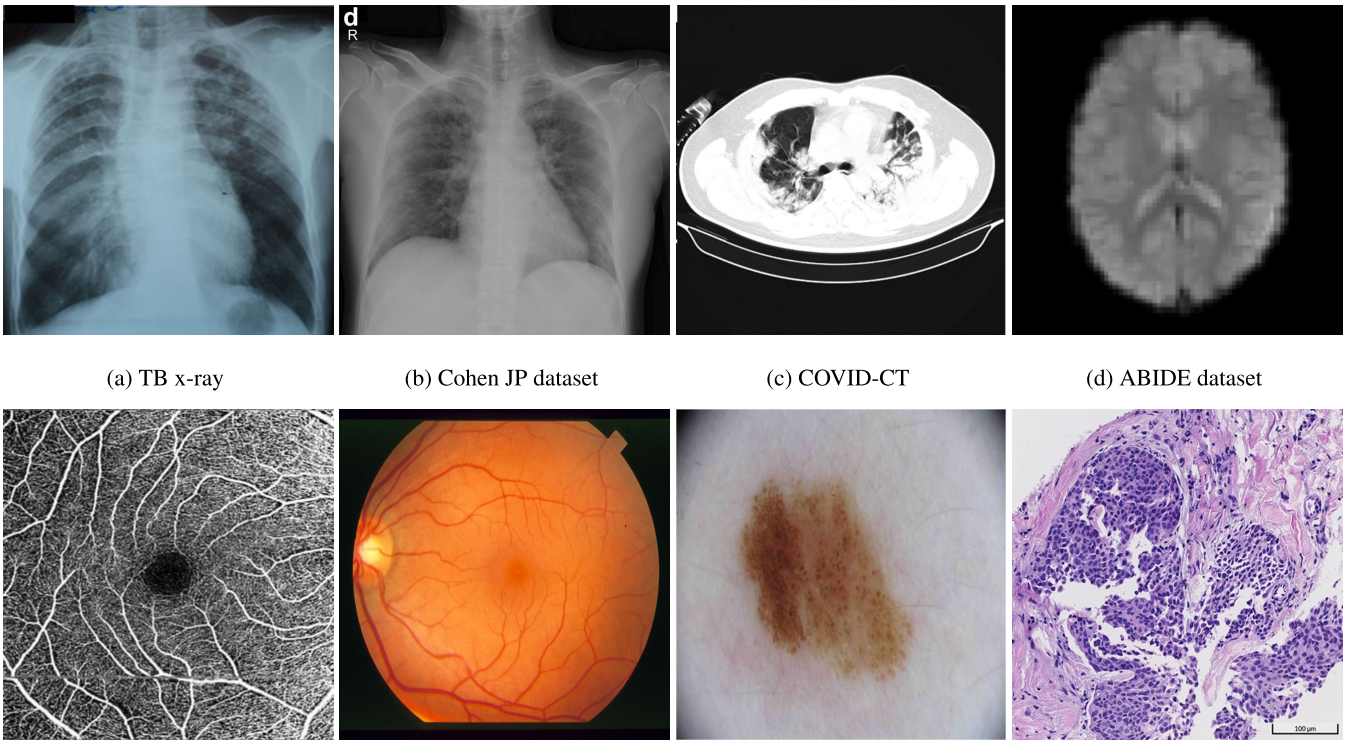
RQ2 What problems were solved?

Almost all of the articles considered in our investigation solved an universal problem which is 'ensure the security of private data'. Data is always a key factor while we need to train a ML model, besides it is a challenge to protect the data from potential security and privacy threats. These threats are more crucial in Electronic Health Record (EHR) data analysis. Sharing EHR data includes patients' private information, above all their identity could be under risk to expose publicly. Similarly in medical image analysis, maintaining privacy of users' data such as X-ray, CT, MRI images is going to be difficult with traditional ML layout. Hence, FL is a privacy preserving way of training AI algorithms, allows to move the model to the data rather than moving the data to the model and this makes it very useful in cases where sensitive data cannot be shared. Since researchers are working for a long time on the application domains that we have discussed in the previous section, now they applied the same fenomena with privacy preserving FL as their experimental research.

B. DATASET

RQ3 What type of dataset used?

We divided the used datasets in the 17 investigated articles into several categories based on image type. Data type varies from model to model, it actually depends on which domain the model will apply; for example, skin images are used to detect skin cancer. Fig. 5 displays eight different types of medical images collected from various datasets used in the research field. **Lung X-ray image:** As we mentioned, severe cases of some diseases affect particular organs of our body, lung is one of them. Literature shows COVID-19 and pneumonia complications include lung damage which is the reason behind using lung X-ray and CT images in such disease detection models. Likewise, as we know smoking



(a) TB x-ray (b) Cohen JP dataset (c) COVID-CT (d) ABIDE dataset

(e) OCT-A data (f) DR dataset (g) MNIST: HAM10000 dataset (h) TCGA cancer data

FIGURE 5. Different types data samples taken from respective dataset. (a) Tuberculosis infected chest X-ray image, (b) COVID-19 positive chest X-ray image, (c) COVID-19 positive CT image, (d) Brain MRI image for autism spectrum disorders identification, (e) Optical coherence tomography angiography (OCT-A) image of eye, (f) Light-sensitive tissue data of retinal blood vessel, (g) Skin dermoscopy image, and (h) Lung tissue image for cancer detection.

is dangerous for health, which particularly affects our lungs and a key reason for lung cancer. According to our investigation, six articles [18], [23], [25], [27], [29], [32] have used chest X-ray images out of 17. The X-ray datasets considered in the articles are Cohen JP, TB x-ray, CheXpert, Mendeley data, COVIDx, Chest X-ray (CXR), and COVID 2019 dataset. Moreover, Zhang et al. [24] proposed a FL oriented COVID-19 detection model where chest X-ray and CT images were considered from three datasets, Qatar-Dhaka data, COVID-CT, and Figure 1 dataset. **Skin image data:** MNIST: HAM10000 is one of the leading datasets used in skin cancer detection research with deep learning techniques. This repository contains 10,015 dermatoscopic images divided into seven different classes. Połap and the groups used the dataset in a series of articles [17], [19], [22] with FL environment. Similar data has been used in [30], which was released under a dataset challenge competition called ISIC 2019 and contains 25,331 dermoscopy images. **Retina image data:** We found two different articles which have applied retinal images for their FL models. In [31], Zhou et al. used a DR dataset consisting of 3,662 images. The images are noted as five different scalability categories, from no DR to extreme. Their goal was to classify the different levels of DR cases. In another article, Lo et al. [16] collected a total of 153 data samples from four different sources. Their deep learning model performed binary classification to define DR and non-DR samples. **Others:** Adnan et al. [28] used tissue image data, more specifically they proposed a privacy guaranteed ML model where lung tissue images were considered to classify cancer. In addition, Li et al. [20] and Linardos et al. [26] both used MRI images for their models, brain and heart MRI data consequently. We have included all of the dataset name with their references for easy access in Table 5.

RQ4 Are the number of data samples sufficient?

In ML research, it is very established that the more data we have for training purposes the better prediction we will get from the models. Also, chances of model overfitting will increase when we have a smaller dataset; so, it is always advisable to use a larger dataset. For our study, we analyzed the 17 articles by the range of data samples used in the respective research papers. First, we will discuss the articles which have used less than 1,000 samples. As Table 5 shows, four papers [16], [18], [20] and [26] used very small amounts of data, their number of elements are 153, 216, 370, and 180 respectively. Since larger dataset belongs to better potentiality of inside analysis, literally 153 data samples are not technically sound. Next, within the 10,000 sample range, seven papers [21], [24], [25], [27], [28], [31], and [32] used data samples between 2,109 and 6,284 and this number is quite good. Finally, we found six papers [17], [19], [22], [23], [29], [30] all of them have used more than 10,000 images

individually. CheXpert, the largest dataset (overall 223,414 CXR image) found under our investigation was considered by Chakravarty et al. [23].

RQ5 Are non-IID data distribution considered?

There are two forms of federated frameworks exist according to the data distribution, IID and non-IID. IID refers to independent and identically distributed. This can be divided into two parts, independence and identical distribution; independence means that the value (data) of an example does not affect the value of the other. This particular scenario is commonly described by a coin flipping experiment, when a coin is flipped, every time the result of both roles does not depend on the other die. Identically distributed means that the probability of any specific outcome is the same, for example every time flipping a coin there is a 50% chance of getting heads and a 50% chance of getting tails and that value does not change while flipping a coin every time. Non-IID technically inverse from both of the sides. While IID data feature distribution is same across clients, the feature distribution is different in non-IID. The problem is quite common in real life, for example, the appearance of the medical image sample using different machines across different hospitals may not align due to different imaging protocols. Therefore, non-IID data settings mean values are dependent on each other and there are overall trends between them. Generally in FL, local models are trained independently where data distribution is hidden to each other and as a result data type and features could be vary client to client [6], [53], this variation makes non-IID data consideration important in FL research. However, in this study, we investigated FL used in medical image analysis. We observed FL data structure is complicated, especially while the local clients' data are significantly different to each other. Our results show only four papers (we did not find sufficient explanation from [26] and [21]) considered non-IID type along with IID data and the rest 13 did not talk about the content. In [18], Feki et al. divided the collected dataset into four parts for clients data, for IID, they used an equal number of images from both sides, client and class. Moreover, for non-IID data they allocated the samples among classes unequally by a ratio of 66% and 44%. Likewise, Adnan et al. [28] performed FL with IID and non-IID data individually where number of samples were different in each client under non-IID scenario.

C. ML FRAMEWORK

RQ6 Which ML algorithms are used to train local models?

Although FL is the leading focused topic of this investigation, ML techniques make the actual difference when it comes to figure out the overall performance of the models. As usual in the FL framework, each client server data is trained by ML algorithms. Since our review is based on medical image data and this image analysis or computer vision task is mostly conducted by CNN oriented deep learning models. However, to answer the question we searched each of the considered articles and found a variety of using built-in

TABLE 6. ML and FL methods for medical image analysis.

Article	ML Architecture/Training Algorithm	Federated algorithm
[16]	VGG19 (classification)	Basic FL
[17]	AlexNet, VGG16, and Inception	Basic FL
[18]	VGG16 and ResNet50	FedAvg
[19]	AlexNet, VGG16, and Inception	Basic FL
[20]	Deep ANN	Fed algorithm
[21]	CNN	FedAvg
[22]	VGG16, and Inception	Basic FL
[23]	ResNet18	Modified FedAvg
[24]	GhostNet, ResNet50, ResNet101	Basic FL
[25]	VGG16 (modified)	Basic FL
[26]	ResNet18	FedAvg
[27]	ResNet18	SecAgg [54]
[28]	MIL	FedAvg
[29]	MobileNet, ResNet18, MoblieNet, and COVIDNet	Basic FL
[30]	Ensemble CNN	Basic FL
[31]	CNN	FedAvg
[32]	Deep learning	Basic FL

CNN models, such as VGG16, Inception, ResNet18, and many more. VGG16 is a widely considered, reliable, and pre-trained model; five out of 17 surveyed papers considered this CNN model. This model is constructed by 16 layers, 13 convolutional and 3 fully connected layers. Likewise, VGG19 is a 19 layers CNN model and used by Lo et al. [16]. Residual Network (ResNet) is also a commonly used algorithm that can be constructed by different numbers of layers, e.g., ResNet18 ([23], [26], [27], [29]), ResNet50 ([18], [24]), ResNet101 ([24]). Other pre-trained CNN models are Inception ([17], [19], [22]), AlexNet ([17], [19]). Besides, CNN associated customised deep learning models have been used in several articles which is listed in Table 6. Li et al. [20] have used multi-layer perceptron (MLP) classifier which was a deep neural network constructed by one input, hidden, and output layers. Adnan et al. [28] performed image segmentation using a supervised learning approach called Multiple-Instance Learning (MIL) to train the local models.

D. FL IMPLEMENTATION

RQ7 Are any additional security methods implemented?

Data privacy and security both are not similar in practice; privacy covers the use (control, access, and regulate) of data, on the other hand, security defines the potential threats of unauthorized access and malicious attacks. FL mainly preserves the privacy concern since trained models of stakeholders are shared instead of sharing data directly. Still, sharing models can be vulnerable while parameters are exchanged between clients and servers and could be a possible threat against system security [28]. Several additional privacy preserving methods have been described in a systematic review article [83]. However, we found few articles that have

considered additional initiatives for security in FL based medical imaging research. Most of the articles (three out of four) have used Differential Privacy (DP), it allows companies to collect information about their users without compromising the privacy of an individual and the ultimate goal is to be able to share information about a dataset with other people without revealing individuals Personally Identifiable Information (PII) from the dataset [9], [84]. Li et al. [20] used two different mechanisms of DP, Gaussian and Laplace. They defined the noise level α which varied from 0.001 to 1. Similarly, Kaissis et al. [27] have applied both techniques and Adnan et al. [28] have used only Gaussian noise in their experiments. In addition, Połap et al. [17] used encryption and blockchain techniques to make their FL model more secure. They proposed three different learning agents where blockchain technique was applied in Data Management Agent (DMA). According to their description, all patients data (images) have to be their unique IDs, once a request arrive to analysis, it will check whether the ID is exist or not into the database, if not then it will create an unique ID and a block to the blockchain, then transfer the ID to the database with the image.

RQ8 What types federated data partitioning are used?

Mainly three categories of FL described in the previous literature based on the training data distributions across the models. Among the three types, Federated Transfer Learning (FTL) and Vertical FL (VFL) are rarely considered in medical research; another one, Horizontal FL (HFL) was used widely. So, in a horizontal partition the client's database holds many different customers but they are collecting all the same type of data on those customers, in other words ''same features, different samples''. In vertical FL, it has different customers in both but there is an overlap of those customers and they are collecting different features, more specifically ''different features, different samples'' [3], [9], [84]. However, in this investigation we focused on the medical image research and found most of the articles were based on HFL. For example, Feki et al. [18] utilized HFL, they used a chest X-ray image dataset where features are same for all clients but samples are different. Interestingly, Kaissis et al. [27] used two different datasets for training and testing their FL models, the fact is both datasets contain X-ray images (same features) and different data. Only two articles we defined as VFL; [24] have taken three datasets, two X-ray and one CT image based. In the article the authors combined the both types of images and used them to train and test models. In [25], the authors used X-ray and ultrasound images for their federated models. X-ray with CT or ultrasound images are technically different, thus their features will be also different and they used various data features in different clients which makes a VFL scenario.

RQ9 What are the federated frameworks used?

Table 6 represents respective deep learning architectures that were used for training their local models (we discussed in RQ6) and next the federated framework which was mainly the aggression approach of the collected local models in the central server. We observed federated mechanisms are executed in two ways, some articles were driven by formerly proposed build-in FL algorithms and others with basic concepts for aggregation. FedAvg (discussed in Section II), which is a commonly used method in federated aggregation, as Table 6 shows six articles considered this algorithm. Likewise, [20] and [27] used two different federated algorithms named Fed, secure aggregation (SecAgg) respectively. SecAgg is a secure model aggregation for FL also proposed by Google in 2016. Połap and Woźniak [19] proposed a meta-heuristic search based federated model, first they calculated average loss of all local models and then selected only models that have scored higher than the average loss for aggregation in server. Mainly all of them pursue fundamental concepts of FL but they implemented it in different ways. However, the described above federated aggregation process has no impact on the model performance, it is all about engineering the data distribution in a decentralized and collaborative manner.

E. EXPERIMENTAL

RQ10 What are the performance measures used in the studies?

The final and startling step of any ML setting is to assess how good the model is through performance evaluation. The basic idea is to develop a ML model using some training samples and test this train model on some other unknown data. However, the training error is not very useful for actual evaluation, because it is easy to overfit the training data by using complex models which do not generalize well to future samples. Contrariwise, testing error is the key metric since it has a better approximation of the true performance of the model on future samples. Thereby, we only considered testing performance throughout our review. As we found from this investigation, classification and segmentation both tasks were used and that is why their performance were also evaluated in different ways. In Fig. 6 we have presented the number of articles using different performance metrics. Most of the experiments (14 out of 17) were evaluated by accuracy. Recall was the second commonly used measurement criteria, considered by five articles. Area Under the Curve (AUC) score three and precision were used two times.

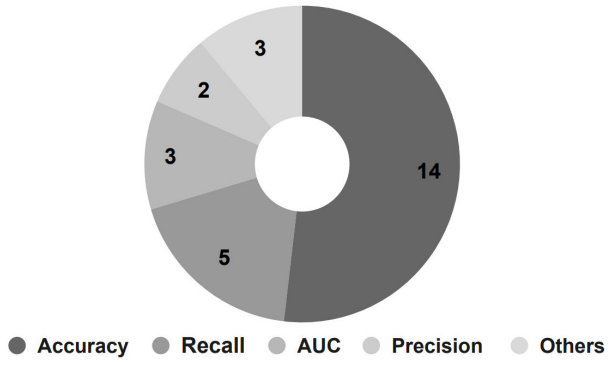


FIGURE 6. Number of articles applied different performance metrics.

RQ11 How is the performance of the FL frameworks reported?

This question is for getting the overview of performance achieved by the FL based models in the 17 articles. Performance assessment is the ultimate part of any ML model where the conducted experiment is evaluated by different matrices. Our investigation revealed 14 articles worked on data classification (binary and multi-class), one article worked on data segmentation, and remaining two considered both of them (listed in Table 4). Usually performance of classification tasks is assessed by accuracy, it represents the report of correctly identified samples from all of the data [14]. We divided the performance into three categories according to the achieved accuracy by the 17 studies: high (>=90%), medium (80%-89%), and low (<80%). Table 7 summarised the performance scores of all articles.

High: We found eight articles have an accuracy of 90% or more. Feki et al. [18] performed binary classification, their accuracy score is highest, for FL+VGG16 with data augmentation model 94.4% and for FL+VGG16 93.57%. Połap and Woźniak [19] used the inception91 classifier for the FL model and obtained an accuracy of 91%. Score of [25], [30], and [24] is not clear, they discussed the accuracy between 90-95%. Article [32] and [27] achieved an accuracy of 90.61% and 90% respectively. Yan et al. [29] presented their results using sensitivity, their highest score was 91.26%.

Medium: In [22], the author classified the images as diseases and not a disease, their proposed VGG based FL model achieved 89.82% accuracy. Lo et al. [16] performed classification and segmentation both tasks on different datasets, the classification and segmentation accuracy for SFU dataset were 88% and 85% respectively, classification accuracy of OHSU dataset was 89%. In [26], Linardos et al. considered AUC, the highest score achieved by the FL model was 89%. Adnan et al. [28] conducted binary classification with an accuracy of 85%.

Low: The rest three articles performed multi-class classification, where Chakravarty et al. [23] 14, Połap et al. [17] seven, and Li et al. [20] have considered four classes, their acquired performances are AUC 80%, accuracy 70% and 76% respectively.

RQ12 How perform the FL approach compared to the conventional models?

Last research question explores the comparative performance analysis between FL and traditional ML image processing research. This query is important while we want to discuss the effect, contribution, and drawback of using FL in medical image analysis. To answer this question we intensively collected experiment results from both areas, 17 FL articles and their relevant conventional models. We already described the performance of the FL models in the previous question and here we will present the results of usual ML models and then the comparative analysis. In Table 7 we summarized the performance of all articles in this review and we presented the results of one or more similar articles opposite to each of the articles to make a comparison chart. To do so, we extensively investigated dozens of research papers that analyzed medical images by traditional ML to explore best matching options which was essential for a reliable comparison. Several conditions were applied in this criteria based on the structural and experimental similarity between ML and FL papers, such as we considered the papers which used similar datasets, algorithms, and performance measures. We expect maintaining this condition will ensure an accurate comparison among the two parties. Our investigation shows in Table 7 that all of the ML models have improved accuracy compared to their respective FL models in existing literature, more specifically we found better ML results against every FL article. For instance, Połap et al. [17] have achieved accuracy with federated VGG16 70% and Inception 67%, however in ML part, Jain et al. [56] achieved 79.23% accuracy with Inception and Liu et al. [57] 87% with ResNet50; all of three have considered the MNIST: HAM10000 dataset. Then as well Chakravarty et al. [23] has an AUC score of 80% with FL environment, but with same dataset and ML algorithm article [67] and [68] have 86% and 87% AUC respectively.

TABLE 7. Performance (accuracy) comparison, the results of each of the reviewed articles and their respective compatible ML models with references.

Article	FL	ML[references]
[16]	89%	92% [55]
[17]	70%	79.23% [56], 87% [57]
[18]	94.4%	99.96% [58], 98% [59]
[19]	91%	93.4% [60], 95% [61]
[20]	Average 76%	84.5% [62], 85.3% [63]
[21]	*ND*	–
[22]	89.82%	97.7% [64], 93.4% [65], 93.4% [66]
[23]	Average AUC: 80%	Average AUC: 86% [67], 87% [68]
[24]	90%-95%	98.28% [69]
[25]	Around 90%	100% [92], 98% [59]
[26]	AUC 89%	AUC 93% [81], Accuracy 99% [82]
[27]	90%	98.8% [70], 96.4% [71], 97.94% [72]
[28]	85%	95% [73]
[29]	91.26%	97.89% [74], 97.83% [75]
[30]	Around 90%	93% [77], 91% [78]
[31]	Around 80%	95% [79], 96.7% [80]
[32]	91.61%	98% [76]

ND - No specific data was found

V. OPEN ISSUES AND CHALLENGES

FL is still a young research field, so it is difficult to draw a remark on the rejection and acceptance. However, here we have discussed the issues and challenges found in the reviewed articles regarding the application of FL in medical image. Generally, FL is invented to fulfill the privacy concern of private data, unfortunately it does not cover all potential privacy threats [93]. However, we described model performance, data heterogeneity, and federated model efficiency issues found from the review below:

A. PRIVACY AND SECURITY

Medical image data is created by personal information of patients and no one can share this data for AI applications without reliable data protection. FL makes the data sharing between the different institutions with some privacy guarantees by an advanced data management and model construction process, all we have described in Section II. FL is different compared to ML models where the training process is exposed to multiple parties, we do not know the motive of every participant, it is an issue of trust among them; so this additional communication increases the risk of leakage data via reverse engineering. Meanwhile, we observed two further privacy measures used in federated medical image processing, differential privacy and secure aggregation. Differential privacy involves adding carefully selected noise to the outputs and can either be done by the individual clients or server level, secure aggregation is a cryptographic technique (e.g., blockchain technology), ensures the server can only see the aggregate of thousands of updates rather than individual model updates. But the reality is every privacy mechanism comes with a significant computational cost on the federation.

B. DATA HETEROGENEITY

Our investigation shows data heterogeneity could occur in two ways: number of samples are different (non-IID data) and data features are different (VFL) among the clients. Usually, the number of produced data in hospitals are not identical and in FL, clients can have different data distributions, this uneven distribution of data of client sides might provide opposing gradient updates to the server which is challenging to tackle. Furthermore, practically features of federated datasets are not the same in many cases, for instance X-ray and CT images data can be used in two different clients which makes trouble during aggregate the models parameters centrally in a FL setting.

C. OVERALL MODEL PERFORMANCE

The first impression of an AI model is the performance, how accurately the model accomplished the task. High performance accuracy makes the model more acceptable than a model that achieved a lower score. We previously discussed the federated model performance and compared them with traditional ML models (RQ12). Our findings show FL failed to perform better than ML with similar model structures, this drawback claims us to reevaluate the usefulness of FL in medical image.

D. FEDERATED ARCHITECTURE

Training a personalized model on each of the clients is not difficult in FL, problems emerge when all of the model output transfers to the central server and passes through an aggregation process. We observed that the federated models presented in the reviewed articles are mostly theoretical and less practically implemented, few articles included their open source code with their articles. Since the research started in the field a couple of years ago, the research method and materials need to be more easily accessible to future researchers. Besides, we usually have a very controlled setting in research, but the question comes when we try to aim for huge datasets to simulate in a real-world scenario.

VI. LIMITATIONS

In this section we have admitted the limitations of this study. First, we searched all prominent databases for article collection where some journals and conference proceedings were with subscription download policy. In some of such cases, we could not grab the papers from the sources. Although, we tried for an alternative way, sent email to the corresponding authors and requested for a full text of the required article. However, still we failed to reach some of them ([94] and [95]) which is limiting the range of this survey. In addition, our inclusion and exclusion process removed articles from the initial fleet and preprint articles were not included there, besides we could not explore all of the searching databases so it could be possible that we missed to include any relevant article(s) on the topic. We did not experiment the models used in the 17 articles under our supervision, for a precise review that would have been more effective. Overall, it is difficult to conclude this study with strong and tested historical evidence, because our review was on very limited time and with insufficient resources since FL was recently introduced.

VII. CONCLUSION AND FUTURE DIRECTIONS

One of the most popular and effective diagnosis methods is imaging techniques in the medical sector. This practice is increasing day by day and produces tons of image data. AI has lots of opportunities in medical imaging using this data, but clinical use of AI and ML is very limited right now. In research direction, creating a publicly shareable image dataset is very difficult for the medical domain. The major hurdle behind data share and collaboration is privacy issues which are less prioritized in typical centralized models. Apart from this concept, federated or distributed learning is different, here a data-driven learning model is shared not the data directly. In this study, we systematically reviewed the articles that considered FL in their ML based medical image research. We elaborately discussed from every perspective, including demographic data, privacy appearance, datasets, FL characteristics, model implementation, and performance comparison. We noticed in one of our previous articles [33] that deep learning oriented COVID-19 detection using X-ray and CT images has high accuracy, most of them achieved more than 95% accuracy. We further observed a similar trend in this study, here COVID-19 detection research articles are the top scorers with FL mechanism. Although, the scores under FL are comparatively lower than general models, as listed in Table 7. Performance of other application domains with FL models were also not mentionable. Besides, previous articles point out the implementation of federated models is relatively complex, it requires extra communication and maintenance trouble. However, it is favorable to become acquainted that

TABLE 8. Quality questions and the scores achived by the 17 articles.

Quality Question\Reference	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32
QQ01 Is the objective of the study clear?	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
QQ02 Is the dataset size sufficient for this type of studies?	-1	1	-1	1	-1	1	1	1	1	1	-1	1	1	1	1	1	1
QQ03 Is the federated data handling procedure clearly defined?	1	0	1	0	1	0	0	0	1	0	0	0	0	0	0	0	0
QQ04 Is non-IID data distribution considered?	-1	-1	1	-1	-1	0	-1	-1	-1	-1	0	-1	1	-1	-1	-1	-1
QQ05 Are any additional privacy methods used?	-1	1	-1	-1	1	-1	-1	-1	-1	-1	-1	1	1	-1	-1	-1	-1
QQ06 Does the author provide sufficient detail about the experiment?	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0
QQ07 Are the learning techniques clearly defined?	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0
QQ08 Are the results clearly stated?	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
QQ09 Is there a comparison among techniques?	-1	-1	0	1	0	-1	1	-1	0	0	0	1	0	0	-1	-1	-1
QQ10 Are the limitations of the study given?	-1	-1	-1	-1	1	-1	1	-1	-1	1	0	1	-1	-1	-1	-1	-1
QQ11 Does the study add value to the existing literature?	0	0	1	1	0	1	1	0	1	1	1	1	1	1	1	0	0
QQ12 Does the study provide any tool or source code online?	1	0	-1	-1	1	-1	-1	-1	-1	-1	1	1	-1	-1	-1	-1	-1
Quality scores	**-1**	**2**	**3**	**3**	**6**	**2**	**5**	**0**	**3**	**4**	**4**	**9**	**6**	**2**	**1**	**0**	**-2**

the research field got lots of attention and publications within a very little time, that is why we can hope for promising progress of FL in medical image analysis in future. At this stage, we have summarized our findings below for future direction to the researchers who are interested to contribute in the field:

- Privacy concern is not fully solved in FL, however, we cannot deny the importance of decentralized concepts. It could be effective for collaborative ML in medical image research, thus researchers should emphasize on the implementation of additional privacy protection in a cost effective way.
- Datasets in the research are collected from various sources and for various purposes where experimental results could differ enormously. There is no particular or benchmark dataset available in federated medical imaging research; need to build some standard datasets to avoid biased data and data heterogeneity problems.
- Similarly no benchmark FL model has been presented yet in this field, such that initiative will assist to build robust AL models for further research.
- In truth, collaborative models data are prone to be heterogeneous, various classes of data are collaborating there. But our results show the accuracy of multi-class classification is very low (as described in RQ11) which needs to be addressed in future research.
- Federated models achieved satisfactory performance in some cases but we cannot narrate as an alternative in the accuracy race with ML models.
- There are many weaknesses observed in current publications (papers investigated in this review) of this field, we included the article quality checklist and results in A. Future research could consider the quality analysis questionnaires for article quality improvement.

No doubt FL is something that might be in the future horizon. But still there are some technical problems, that challenges need to be tackled before FL is going to be applied vastly. Best of our knowledge this is the first SLR and we believe this review is a reflection of FL research in the area of medical imaging.

APPENDIX
QUALITY ANALYSIS

Table 8 shows 12 Quality Questions (QQ) and scores, mostly motivated from our previous article [14]. The goal of such inquiry was to check the basic quality of the articles published in FL oriented medical imaging. However, each question has one score for one article and a total score of 12 for an individual. We considered the QQ answer in three forms of

scoring, "Yes (1)", "Partially Yes (0)", and "No (−1)". The article which clearly supports the question is Yes, partially supported or where no clear answer found is Partially Yes, and lastly fully disagreed is No. We investigated each of the articles to find the answer and assigned the scores in respective columns. As the table interprets most of the articles have failed to fulfill the quality requirement. Highest score is 9 out of 12 gained by [27], followed by six for [20] and [28] both articles individually. The score indicates in some areas quality has been maintained poorly in the research papers, a reason could be that lots of attention made a rush on FL research among the contributors.

REFERENCES

[1] A. Maier, S. Steidl, V. Christlein, and J. Hornegger, *Medical Imaging Systems: An Introductory Guide*. Cham, Switzerland: Springer, 2018.
[2] Z. Zhang and E. Sejdic, "Radiological images and machine learning: Trends, perspectives, and prospects," *Comput. Biol. Med.*, vol. 108, pp. 354–370, May 2019.
[3] A. Rauniyar, D. Haileselassie Hagos, D. Jha, J. E. Håkegård, U. Bagci, D. B. Rawat, and V. Vlassov, "Federated learning for medical applications: A taxonomy, current trends, challenges, and future research directions," 2022, *arXiv:2208.03392*.
[4] H. B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. Y. Arcas, "Federated learning of Deep Networks using model averaging," 2016, *arXiv:1602.05629*.
[5] B. Pfitzner, N. Steckhan, and B. Arnrich, "Federated learning in a medical context: A systematic literature review," *ACM Trans. Internet Technol.*, vol. 21, no. 2, pp. 1–31, Jun. 2021.
[6] Prayitno, C.-R. Shyu, K. T. Putra, H.-C. Chen, Y.-Y. Tsai, K. S. Hossain, W. Jiang, and Z.-Y. Shae, "A systematic review of federated learning in the healthcare area: From the perspective of data properties and applications," *Appl. Sci.*, vol. 11, no. 23, p. 11191, Nov. 2021.
[7] M. G. Crowson, D. Moukheiber, A. R. Arévalo, B. D. Lam, S. Mantena, A. Rana, D. Goss, D. W. Bates, and L. A. Celi, "A systematic review of federated learning applications for biomedical data," *PLOS Digit. Health*, vol. 1, no. 5, May 2022, Art. no. e0000033.
[8] A. Chowdhury, H. Kassem, N. Padoy, R. Umeton, and A. Karargyris, "A review of medical federated learning: Applications in oncology and cancer research," in *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*. Springer, Jul. 2022, doi: 10.1007/978-3-031-08999-2_1.
[9] D. C. Nguyen, Q.-V. Pham, P. N. Pathirana, M. Ding, A. Seneviratne, Z. Lin, O. Dobre, and W.-J. Hwang, "Federated learning for smart healthcare: A survey," *ACM Comput. Surv.*, vol. 55, no. 3, pp. 1–37, Apr. 2023.
[10] A. Naeem, T. Anees, R. A. Naqvi, and W.-K. Loh, "A comprehensive analysis of recent deep and federated-learning-based methodologies for brain tumor diagnosis," *J. Personalized Med.*, vol. 12, no. 2, p. 275, Feb. 2022.
[11] J. Xu, B. S. Glicksberg, C. Su, P. Walker, and J. Bian, "Federated learning for healthcare informatics," *J. Healthc Inform. Res.*, vol. 5, pp. 1–19, Dec. 2021.
[12] N. Rieke, J. Hancox, W. Li, F. Milletarì, H. R. Roth, S. Albarqouni, S. Bakas, M. N. Galtier, B. A. Landman, K. Maier-Hein, S. Ourselin, M. Sheller, R. M. Summers, A. Trask, D. Xu, M. Baust, and M. J. Cardoso, "The future of digital health with federated learning," *npj Digit. Med.*, vol. 3, no. 1, p. 119, Sep. 2020.
[13] H. B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. Y. Arcas, "Communication-efficient learning of deep networks from decentralized data," 2016, *arXiv:1602.05629*.
[14] M. F. Sohan and A. Basalamah, "A systematic literature review and quality analysis of Javascript malware detection," *IEEE Access*, vol. 8, pp. 190539–190552, 2020.
[15] D. Moher, A. Liberati, J. Tetzlaff, and D. G. Altman, "Preferred reporting items for systematic reviews and meta-analyses: The Prisma statement," *PLoS Med.*, vol. 6, no. 7, p. 264, 2009.
[16] J. Lo, T. T. Yu, D. Ma, P. Zang, J. P. Owen, Q. Zhang, R. K. Wang, M. F. Beg, A. Y. Lee, Y. Jia, and M. V. Sarunic, "Federated learning for microvasculature segmentation and diabetic retinopathy classification of OCT data," *Ophthalmol. Sci.*, vol. 1, no. 4, Dec. 2021, Art. no. 100069.
[17] D. Połap, G. Srivastava, and K. Yu, "Agent architecture of an intelligent medical system based on federated learning and blockchain technology," *J. Inf. Secur. Appl.*, vol. 58, May 2021, Art. no. 102748.
[18] I. Feki, S. Ammar, Y. Kessentini, and K. Muhammad, "Federated learning for COVID-19 screening from chest X-ray images," *Appl. Soft Comput.*, vol. 106, Jul. 2021, Art. no. 107330.
[19] D. Połap and M. Wozniak, "Meta-heuristic as manager in federated learning approaches for image processing purposes," *Appl. Soft Comput.*, vol. 113, Dec. 2021, Art. no. 107872.
[20] X. Li, Y. Gu, N. Dvornek, L. H. Staib, P. Ventola, and J. S. Duncan, "Multi-site fMRI analysis using privacy-preserving federated learning and domain adaptation: ABIDE results," *Med. Image Anal.*, vol. 65, Oct. 2020, Art. no. 101765.
[21] D. Yang, Z. Xu, W. Li, A. Myronenko, H. R. Roth, S. Harmon, S. Xu, B. Turkbey, E. Turkbey, X. Wang, W. Zhu, G. Carrafiello, F. Patella, M. Cariati, H. Obinata, H. Mori, K. Tamura, P. An, B. J. Wood, and D. Xu, "Federated semi-supervised learning for COVID region segmentation in chest CT using multi-national data from China, Italy, Japan," *Med. Image Anal.*, vol. 70, May 2021, Art. no. 101992.
[22] D. Poap, "Fuzzy consensus with federated learning method in medical systems," *IEEE Access*, vol. 9, pp. 150383–150392, 2021.
[23] A. Chakravarty, A. Kar, R. Sethuraman, and D. Sheet, "Federated learning for site aware chest radiograph screening," in *Proc. IEEE 18th Int. Symp. Biomed. Imag. (ISBI)*, Apr. 2021, pp. 1077–1081.
[24] W. Zhang, T. Zhou, Q. Lu, X. Wang, C. Zhu, H. Sun, Z. Wang, S. K. Lo, and F.-Y. Wang, "Dynamic-fusion-based federated learning for COVID-19 detection," *IEEE Internet Things J.*, vol. 8, no. 21, pp. 15884–15891, Nov. 2021.
[25] A. Qayyum, K. Ahmad, M. A. Ahsan, A. Al-Fuqaha, and J. Qadir, "Collaborative federated learning for healthcare: Multi-modal COVID-19 diagnosis at the edge," *IEEE Open J. Comput. Soc.*, vol. 3, pp. 172–184, 2022.
[26] A. Linardos, K. Kushibar, S. Walsh, P. Gkontra, and K. Lekadir, "Federated learning for multi-center imaging diagnostics: A simulation study in cardiovascular disease," *Sci. Rep.*, vol. 12, no. 1, p. 3551, Mar. 2022.
[27] G. Kaissis, A. Ziller, J. Passerat-Palmbach, T. Ryffel, D. Usynin, A. Trask, I. Lima, J. Mancuso, F. Jungmann, M.-M. Steinborn, A. Saleh, M. Makowski, D. Rueckert, and R. Braren, "End-to-end privacy preserving deep learning on multi-institutional medical imaging," *Nature Mach. Intell.*, vol. 3, pp. 473–484, Jun. 2021.
[28] M. Adnan, S. Kalra, J. C. Cresswell, G. W. Taylor, and H. R. Tizhoosh, "Federated learning and differential privacy for medical image analysis," *Sci. Rep.*, vol. 12, no. 1, p. 1953, Feb. 2022.
[29] B. Yan, J. Wang, J. Cheng, Y. Zhou, Y. Zhang, Y. Yang, L. Liu, H. Zhao, C. Wang, and B. Liu, "Experiments of federated learning for COVID-19 chest X-ray images," *Advances in Artificial Intelligence and Security*. Springer, 2021, pp. 41–53.
[30] M. A. Hashmani, S. M. Jameel, S. S. H. Rizvi, and S. Shukla, "An adaptive federated machine learning-based intelligent system for skin disease detection: A step toward an intelligent dermoscopy device," *Appl. Sci.*, vol. 11, no. 5, p. 2145, Feb. 2021.
[31] S. Zhou, B. Landman, Y. Huo, and A. Gokhale, "Communication-efficient federated learning for multi-institutional medical image classification," in *Proc. SPIE*, 2022, pp. 6–12.
[32] M. A. Salam, S. Taha, and M. Ramadan, "COVID-19 detection using federated machine learning," *PLoS ONE*, vol. 16, no. 6, Jun. 2021, Art. no. e0252573.
[33] M. F. Sohan, A. Basalamah, and M. Solaiman, "COVID-19 detection using machine learning: A large scale assessment of X-ray and CT image datasets," *J. Electron. Imag.*, vol. 31, no. 4, Mar. 2022, Art. no. 041212.
[34] K. Munir, H. Elahi, A. Ayub, F. Frezza, and A. Rizzi, "Cancer diagnosis using deep learning: A bibliographic review," *Cancers*, vol. 11, no. 9, p. 1235, Aug. 2019.
[35] M. Heisler, F. Chan, Z. Mammo, C. Balaratnasingam, P. Prentasic, G. Docherty, M. Ju, S. Rajapakse, S. Lee, A. Merkur, A. Kirker, D. Albiani, D. Maberley, K. Bailey Freund, M. Faisal Beg, S. Loncaric, M. V. Sarunic, and E. V. Navajas, "Deep learning vessel segmentation and quantification of the foveal avascular zone using commercial and prototype OCT—A platforms," 2019, *arXiv:1909.11289*.
[36] P. Tschandl, C. Rosendahl, and H. Kittler, "The HAM 10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions," *Sci. Data*, vol. 5, no. 1, Aug. 2018.

[37] J. P. Cohen, P. Morrison, L. Dao, K. Roth, T. Q. Duong, and M. Ghassemi, "COVID-19 image data collection: Prospective predictions are the future," 2020, *arXiv:2006.11988*.

[38] S. Jaeger, S. Candemir, S. Antani, Y.-X. J. Wáng, P.-X. Lu, and G. Thoma, "Two public chest X-ray datasets for computer-aided screening of pulmonary diseases," *Quant. Imag. Med. Surg.*, vol. 4, no. 6, pp. 475–477, Nov. 2014, Accessed: Oct. 18, 2022. [Online]. Available: https://qims.amegroups.com/article/view/5132/6030

[39] A. Di Martino et al., "The autism brain imaging data exchange: Towards a large-scale evaluation of the intrinsic brain architecture in autism," *Mol. Psychiatry*, vol. 19, no. 6, pp. 659–667, Jun. 2013.

[40] J. Irvin, P. Rajpurkar, M. Ko, Y. Yu, S. Ciurea-Ilcus, C. Chute, H. Marklund, B. Haghgoo, R. Ball, K. Shpanskaya, J. Seekins, D. A. Mong, S. S. Halabi, J. K. Sandberg, R. Jones, D. B. Larson, C. P. Langlotz, B. N. Patel, M. P. Lungren, and A. Y. Ng, "CheXpert: A large chest radiograph dataset with uncertainty labels and expert comparison," in *Proc. AAAI Conf. Artif. Intell.*, 2019, pp. 590–597, Accessed: Oct. 18, 2022. [Online]. Available: https://ojs.aaai.org/index.php/AAAI/article/view/3834

[41] M. E. Chowdhury, T. Rahman, A. Khandakar, R. Mazhar, M. A. Kadir, Z. B. Mahbub, K. R. Islam, M. S. Khan, A. Iqbal, N. A. Emadi, M. B. Reaz, and M. T. Islam, "Can ai help in screening viral and COVID-19 pneumonia?" *IEEE Access*, vol. 8, pp. 132665–132676, 2020.

[42] X. Yang, X. He, J. Zhao, Y. Zhang, S. Zhang, and P. Xie, "COVID-CT-dataset: A CT scan dataset about COVID-19," 2020, *arXiv:2003.13865*.

[43] Agchung. *Agchung/Figure1-COVID-Chestxray-Dataset: Figure 1 COVID-19 Chest X-Ray Dataset Initiative*. GitHub. Accessed: Oct. 18, 2022. [Online]. Available: https://github.com/agchung/Figure1-COVID-chestxray-dataset

[44] V. M. Campello et al., "Multi-centre, multi-vendor and multi-disease cardiac segmentation: The M&Ms challenge," *IEEE Trans. Med. Imag.*, vol. 40, no. 12, pp. 3543–3554, Dec. 2021.

[45] O. Bernard et al., "Deep learning techniques for automatic MRI cardiac multi-structures segmentation and diagnosis: Is the problem solved?" *IEEE Trans. Med. Imag.*, vol. 37, no. 11, pp. 2514–2525, Nov. 2018.

[46] D. Kermany, K. Zhang, and M. Goldbaum. (Jan. 1, 2018). *Large Dataset of Labeled Optical Coherence Tomography (OCT) and Chest X-Ray Images*. Mendeley Data. Accessed: Oct. 18, 2022. [Online]. Available: https://data.mendeley.com/datasets/rscbjbr9sj/3

[47] Project-MONAI. *Project-Monai/Monai: AI Toolkit for Healthcare Imaging*. GitHub. Accessed: Oct. 18, 2022. [Online]. Available: https://github.com/Project-MONAI/MONAI/

[48] *Medical Segmentation Decathlon*. Accessed: Oct. 18, 2022. [Online]. Available: http://medicaldecathlon.com/

[49] K. Tomczak, P. Czerwinska, and M. Wiznerowicz, "Review the cancer genome atlas (TCGA): An immeasurable source of knowledge," *Współczesna Onkologia*, vol. 1A, pp. 68–77, Jan. 2015.

[50] L. Wang, Z. Q. Lin, and A. Wong, "COVID-Net: A tailored deep convolutional neural network design for detection of COVID-19 cases from chest X-ray images," *Sci. Rep.*, vol. 10, no. 1, pp. 1–12, Nov. 2020.

[51] *ISIC Challenge*. Accessed: Oct. 18, 2022. [Online]. Available: https://challenge.isic-archive.com/landing/2019/

[52] J. Y. Choi, T. K. Yoo, J. G. Seo, J. Kwak, T. T. Um, and T. H. Rim, "Multi-categorical deep learning neural network to classify retinal images: A pilot study employing small database," *PLoS ONE*, vol. 12, no. 11, Nov. 2017, Art. no. e0187336.

[53] H. Zhu, J. Xu, S. Liu, and Y. Jin, "Federated learning on non-IID Data: A survey," *Neurocomputing*, vol. 465, pp. 371–390, Sep. 2021, Accessed: Oct. 18, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/abs/pii/S0925231221013254

[54] K. Bonawitz, V. Ivanov, B. Kreuter, A. Marcedone, H. B. McMahan, S. Patel, D. Ramage, A. Segal, and K. Seth, "Practical secure aggregation for federated learning on user-held data," 2016, *arXiv:1611.04482*.

[55] M. Heisler, S. Karst, J. Lo, Z. Mammo, T. Yu, S. Warner, D. Maberley, M. F. Beg, E. V. Navajas, and M. V. Sarunic, "Ensemble deep learning for diabetic retinopathy detection using optical coherence tomography angiography," *Transl. Vis. Sci. Technol.*, vol. 9, no. 2, p. 20, Apr. 2020.

[56] S. Jain, U. Singhania, B. Tripathy, E. A. Nasr, M. K. Aboudaif, and A. K. Kamrani, "Deep learning-based transfer learning for classification of skin cancer," *Sensors*, vol. 21, no. 23, p. 8142, Dec. 2021.

[57] Y. Liu, Z. Wang, Z. Li, J. Li, T. Li, P. Chen, and R. Liang, "Multiscale ensemble of convolutional neural networks for skin lesion classification," *IET Image Process.*, vol. 15, no. 10, pp. 2309–2318, Aug. 2021.

[58] D. Das, K. C. Santosh, and U. Pal, "Truncated inception net: COVID-19 outbreak screening using chest X-rays," *Phys. Eng. Sci. Med.*, vol. 43, no. 3, pp. 915–925, Sep. 2020.

[59] N. S. Punn and S. Agarwal, "Automated diagnosis of COVID-19 with limited posteroanterior chest X-ray images using fine-tuned deep neural networks," *Int. J. Speech Technol.*, vol. 51, no. 5, pp. 2689–2702, May 2021.

[60] G. Chowdhary, N. K. Toppo, and D. Das, "Skin lesion diagnosis in healthcare-cyber physical system," in *Proc. IEEE Int. Conf. Innov. Technol. (INOCON)*, Nov. 2020, pp. 1–6.

[61] M. Arshad, M. A. Khan, U. Tariq, A. Armghan, F. Alenezi, M. Younus Javed, S. M. Aslam, and S. Kadry, "A computer-aided diagnosis system using deep learning for multiclass skin lesion classification," *Comput. Intell. Neurosci.*, vol. 2021, pp. 1–15, Dec. 2021.

[62] M. R. Ahmed, Y. Zhang, Y. Liu, and H. Liao, "Single volume image generator and deep learning-based ASD classification," *IEEE J. Biomed. Health Informat.*, vol. 24, no. 11, pp. 3044–3054, Nov. 2020.

[63] X. Li, N. C. Dvornek, J. Zhuang, P. Ventola, and J. S. Duncan, "Brain biomarker interpretation in ASD using deep learning and fMRI," in *Medical Image Computing and Computer Assisted Intervention—MICCAI*. Springer, 2018, pp. 206–214.

[64] M. Kumar, M. Alshehri, R. AlGhamdi, P. Sharma, and V. Deep, "A DE-ANN inspired skin cancer detection approach using fuzzy C-means clustering," *Mobile Netw. Appl.*, vol. 25, no. 4, pp. 1319–1329, Aug. 2020.

[65] F. Afza, M. Sharif, M. A. Khan, U. Tariq, H.-S. Yong, and J. Cha, "Multiclass skin lesion classification using hybrid deep features selection and extreme learning machine," *Sensors*, vol. 22, no. 3, p. 799, Jan. 2022.

[66] U. Bhimavarapu and G. Battineni, "Skin lesion analysis for melanoma detection using the novel deep learning model fuzzy GC-SCNN," *Healthcare*, vol. 10, no. 5, p. 962, May 2022.

[67] L. Seyyed-Kalantari, G. Liu, M. McDermott, I. Y. Chen, and M. Ghassemi, "CheXclusion: Fairness gaps in deep chest X-ray classifiers," in *Proc. Biocomputing*, vol. 2021, 2020, pp. 232–243.

[68] A. Mitra, A. Chakravarty, N. Ghosh, T. Sarkar, R. Sethuraman, and D. Sheet, "A systematic search over deep convolutional neural network architectures for screening chest radiographs," in *Proc. 42nd Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Jul. 2020, pp. 1225–1228.

[69] S. Thakur and A. Kumar, "X-ray and CT-scan-based automated detection and classification of covid-19 using convolutional neural networks (CNN)," *Biomed. Signal Process. Control*, vol. 69, Aug. 2021, Art. no. 102920.

[70] R. Kundu, R. Das, Z. W. Geem, G.-T. Han, and R. Sarkar, "Pneumonia detection in chest X-ray images using an ensemble of deep learning models," *PLoS ONE*, vol. 16, no. 9, Sep. 2021, Art. no. e0256630.

[71] V. Chouhan, S. K. Singh, A. Khamparia, D. Gupta, P. Tiwari, C. Moreira, R. Damasevicius, and V. H. C. de Albuquerque, "A novel transfer learning based approach for pneumonia detection in chest X-ray images," *Appl. Sci.*, vol. 10, no. 2, p. 559, Jan. 2020.

[72] N. Dey, Y.-D. Zhang, V. Rajinikanth, R. Pugalenthi, and N. S. M. Raja, "Customized VGG19 architecture for pneumonia detection in chest X-rays," *Pattern Recognit. Lett.*, vol. 143, pp. 67–74, Mar. 2021.

[73] L. Girard, J. Rodriguez-Canales, C. Behrens, D. M. Thompson, I. W. Botros, H. Tang, Y. Xie, N. Rekhtman, W. D. Travis, I. I. Wistuba, J. D. Minna, and A. F. Gazdar, "An expression signature as an aid to the histologic classification of non–small cell lung cancer," *Clin. Cancer Res.*, vol. 22, no. 19, pp. 4880–4889, 2016.

[74] S. Dong, Q. Yang, Y. Fu, M. Tian, and C. Zhuo, "RCoNet: Deformable mutual information maximization and high-order uncertainty-aware learning for robust COVID-19 detection," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 8, pp. 3401–3411, Aug. 2021.

[75] A. Bar-El, D. Cohen, N. Cahan, and H. Greenspan, "Improved cycle-gan with application to COVID-19 classification," in *Proc. SPIE*, 2021, pp. 296–305.

[76] F. Ucar and D. Korkmaz, "COVIDiagnosis-Net: Deep bayes-squeezenet based diagnosis of the coronavirus disease 2019 (COVID-19) from X-ray images," *Med. Hypotheses*, vol. 140, Jul. 2020, Art. no. 109761.

[77] H. El-Khatib, D. Popescu, and L. Ichim, "Deep Learning–Based methods for automatic diagnosis of skin lesions," *Sensors*, vol. 20, no. 6, p. 1753, Mar. 2020.

[78] H. Nahata and S. P. Singh, "Deep learning solutions for skin cancer detection and diagnosis," in *Learning and Analytics in Intelligent Systems*. Springer, 2020, pp. 159–182.

[79] H. Fu, Y. Xu, S. Lin, D. W. Kee Wong, and J. Liu, "Deepvessel: Retinal vessel segmentation via deep learning and conditional random field," in *Medical Image Computing and Computer-Assisted Intervention—MICCAI*. Springer, 2016, pp. 132–139.
[80] S. S. M. Sheet, T.-S. Tan, M. A. As'ari, W. H. W. Hitam, and J. S. Y. Sia, "Retinal disease identification using upgraded CLAHE filter and transfer convolution neural network," *ICT Exp.*, vol. 8, no. 1, pp. 142–150, Mar. 2022.
[81] C. Luo, C. Shi, X. Li, and D. Gao, "Cardiac MR segmentation based on sequence propagation by deep learning," *PLoS ONE*, vol. 15, no. 4, Apr. 2020, Art. no. e0230415.
[82] S. Tripathi, T. S. Sharan, S. Sharma, and N. Sharma, "An augmented deep learning network with noise suppression feature for efficient segmentation of magnetic resonance images," *IETE Tech. Rev.*, vol. 39, no. 4, pp. 1–14, 2021.
[83] L. Witt, M. Heyer, K. Toyoda, W. Samek, and D. Li, "Decentral and incentivized federated learning frameworks: A systematic literature review," 2022, *arXiv:2205.07855*.
[84] T. R. Gadekallu, Q.-V. Pham, T. Huynh-The, S. Bhattacharya, P. K. R. Maddikunta, and M. Liyanage, "Federated learning for big data: A survey on opportunities, applications, and future directions," 2021, *arXiv:2110.04160*.
[85] H. R. Roth et al., "Federated learning for breast density classification: A real-world implementation," in *Domain Adaptation and Representation Transfer, and Distributed and Collaborative Learning*. Springer, 2020, pp. 181–191.
[86] Q. Dou et al., "Federated deep learning for detecting COVID-19 lung abnormalities in CT: A privacy-preserving multinational validation study," *npj Digit. Med.*, vol. 4, no. 1, p. 60, Mar. 2021.
[87] N. N. Thilakarathne, G. Muneeswari, V. Parthasarathy, F. Alassery, H. Hamam, R. Kumar Mahendran, and M. Shafiq, "Federated learning for privacy-preserved medical Internet of Things," *Intell. Autom. Soft Comput.*, vol. 33, no. 1, pp. 157–172, 2022.
[88] J. Born, N. Wiedemann, M. Cossio, C. Buhre, G. Brändle, K. Leidermann, J. Goulet, A. Aujayeb, M. Moor, B. Rieck, and K. Borgwardt, "Accelerating detection of lung pathologies with explainable ultrasound image analysis," *Appl. Sci.*, vol. 11, no. 2, p. 672, 2021.
[89] P. Patel. (Sep. 17, 2020). *Chest X-ray (COVID-19 & Pneumonia)*. Kaggle. Accessed: Oct. 30, 2022. [Online]. Available: https://www.kaggle.com/datasets/prashant268/chest-xray-covid19-pneumonia
[90] Srk. (Jun. 24, 2021). *Novel Corona Virus 2019 Dataset*. Kaggle. Accessed: Oct. 30, 2022. [Online]. Available: https://www.kaggle.com/datasets/sudalairajkumar/novel-corona-virus-2019-dataset
[91] F. Sattler, K.-R. Müller, and W. Samek, "Clustered federated learning: Model-agnostic distributed multitask optimization under privacy constraints," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 8, pp. 3710–3722, Aug. 2021.
[92] P. R. Bassi and R. Attux, "A deep convolutional neural network for COVID-19 detection using chest X-rays," *Res. Biomed. Eng.*, vol. 38, no. 1, pp. 139–148, Mar. 2022.
[93] K. M. J. Rahman, F. Ahmed, N. Akhter, M. Hasan, R. Amin, K. E. Aziz, A. K. M. M. Islam, M. S. H. Mukta, and A. K. M. N. Islam, "Challenges, applications and design aspects of federated learning: A survey," *IEEE Access*, vol. 9, pp. 124682–124700, 2021.
[94] K. Guo, T. Chen, S. Ren, N. Li, M. Hu, and J. Kang, "Federated learning empowered real-time medical data processing method for smart healthcare," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, early access, Jun. 23, 2022, doi: 10.1109/TCBB.2022.3185395.
[95] S. Sakib, M. M. Fouda, Z. Md Fadlullah, and N. Nasser, "On COVID-19 prediction using asynchronous federated learning-based agile radiograph screening booths," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Jun. 2021, pp. 1–6.

MD FAHIMUZZMAN SOHAN received the B.Sc. degree in software engineering from Daffodil International University, Bangladesh, in 2019. He has published several papers in reputed journals and conferences. His research interests include machine learning, computer vision, and image processing.

ANAS BASALAMAH received the M.Sc. and Ph.D. degrees from Waseda University, Tokyo, in 2006 and 2009, respectively. He was a Postdoctoral Researcher with The University of Tokyo and the University of Minnesota, in 2010 and 2011, respectively. He is currently an Associate Professor with the Department of Computer Engineering, Umm Al-Qura University. His research interests include embedded networked sensing, smart cities, ubiquitous computing, participatory, and urban sensing.