Healthcare Analytics 4 (2023) 100204

Contents lists available at ScienceDirect

Healthcare Analytics

journal homepage: www.elsevier.com/locate/health

A federated learning framework for pneumonia image detection using distributed data

Amer Kareem [a,*], Haiming Liu [b], Vladan Velisavljevic [a]

[a] University of Bedfordshire, Vicarage St, Luton LU1 3JU, United Kingdom
[b] University of Southampton, University Rd, Southampton SO17 1BJ, United Kingdom

ARTICLE INFO

Keywords:
Machine learning techniques
Pneumonia detection
Federated learning
Chest X-ray
Transfer learning
Medical image privacy

ABSTRACT

Pneumonia is one of the serious diseases affecting the lungs. Yearly, over four million people die on average. Therefore, it is essential to have an effective system for early diagnoses. State-of-the-art computer-aided Machine Learning (ML) techniques have been used for pneumonia detection. However, pneumonia X-ray images are visually heterogeneous and complex in pattern recognition. Therefore, a vast amount of dataset is required for effective ML model training. The larger data volume can be collected using the real-time dataset from hospitals and medical institutions. However, due to General Data Protection Regulation (GDPR) and the Data Protection Act (DPA), data sharing is not allowed by the third party. This study is inspired by using real-time datasets in a privacy-preserving fashion while using the framework of federated learning (FL). We have performed experiments using state-of-the-art ML models for medical image classification, including pre-trained Convolutional Neural Network (CNN) models of Alexnet, DenseNet, Residual Neural Network-50 (ResNet50), Inception, and Visual Geometry Group-19 (VGG19). The experiments are performed individually on the models and the FL framework. We compared the results using the evaluation metrics and Area Under the Curve (AUC). The preliminary results show the ResNet-50 stands out in performance on the testing dataset producing an accuracy of 93% significantly.

1. Introduction

According to report [1], approximately 200 million people suffered from pneumonia every year. This report shows that it is more common in young children under the age of 5. The report shows that the death rate of children caused by pneumonia is 16 times higher than caused by cancer and 10 times more than Human immunodeficiency viruses (HIV). The below report analysis provided by World health record shows the causes of death in children under the age of 5 [2]:

A report was issued on World pneumonia day that has predicted the death rate rise to 11 million a year by 2030 [3]. Pneumonia was one of the major causes of death over the last few decades. The advancement in the medical industry has helped in early diagnosis of the disease [4]. The literature analysis shows that the medical specialist adopts multiple methods for pneumonia diagnoses that involves clinical examination, disease symptoms detection and medical records. With the increased usage of computer aided technologies, chest X-ray (CXR) has become popular technique for early diagnosis of pneumonia (see Figs. 1–18).

The related work for pneumonia detection shows that the artificial intelligence (AI) technologies have been extensively adopted recently. It has been considered an effective solution for medical image classification that includes CXR, ultrasound, and other sources. Many areas of AI have been explored and deep learning algorithms have shown effective performance. Our research is inspired by using the AI methods for pneumonia image detection while utilizing the real time data. The next section illustrates scope and motivation of our research.

1.1. Scope and motivation

Medical image classification from X-ray images is a complex task due to visual heterogeneity and high required spatial resolution of the images. In addition, the research using real-time data is challenging because of the limitations imposed by general data protection and regulation (GDPR) and Data Protection Act (DPA). Also, the class imbalance in the dataset is challenging. To perform the effective medical image classification, it requires vast amount of data. Therefore, our solution is to use the real time dataset from multiple hospital and medical institutions, while following GDPR and DPA. Our research is inspired to use the machine learning models to train the dataset from different hospital and medical institutions in a privacy preserving manner.

We have used the federated learning framework which will ensure the data privacy due to its semi-decentralized architecture. On the other hand, we have used machine learning models due to its learning capabilities and effective classification on the different datasets. Our

* Corresponding author.
E-mail addresses: Amer.kareem@study.beds.ac.uk (A. Kareem), H.liu@soton.ac.uk (H. Liu), Vladan.velisavljevic@beds.ac.uk (V. Velisavljevic).

https://doi.org/10.1016/j.health.2023.100204
Received 12 April 2023; Received in revised form 17 May 2023; Accepted 27 May 2023

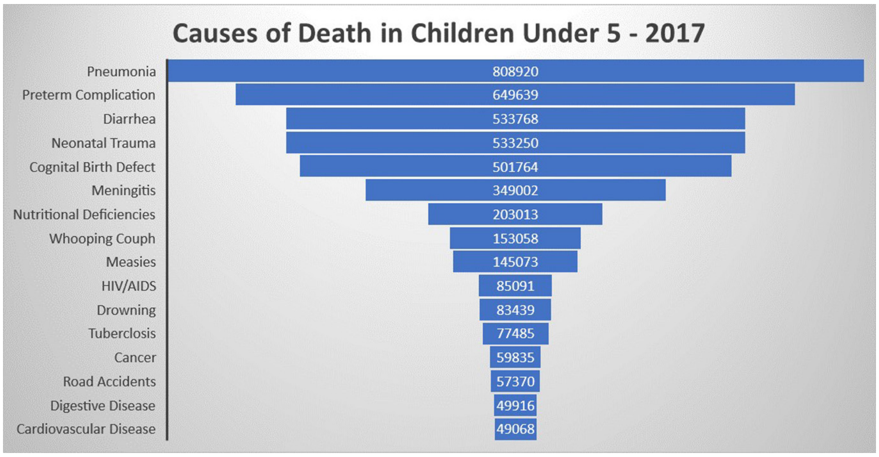

Fig. 1. Causes of death in children under 5 -2017.

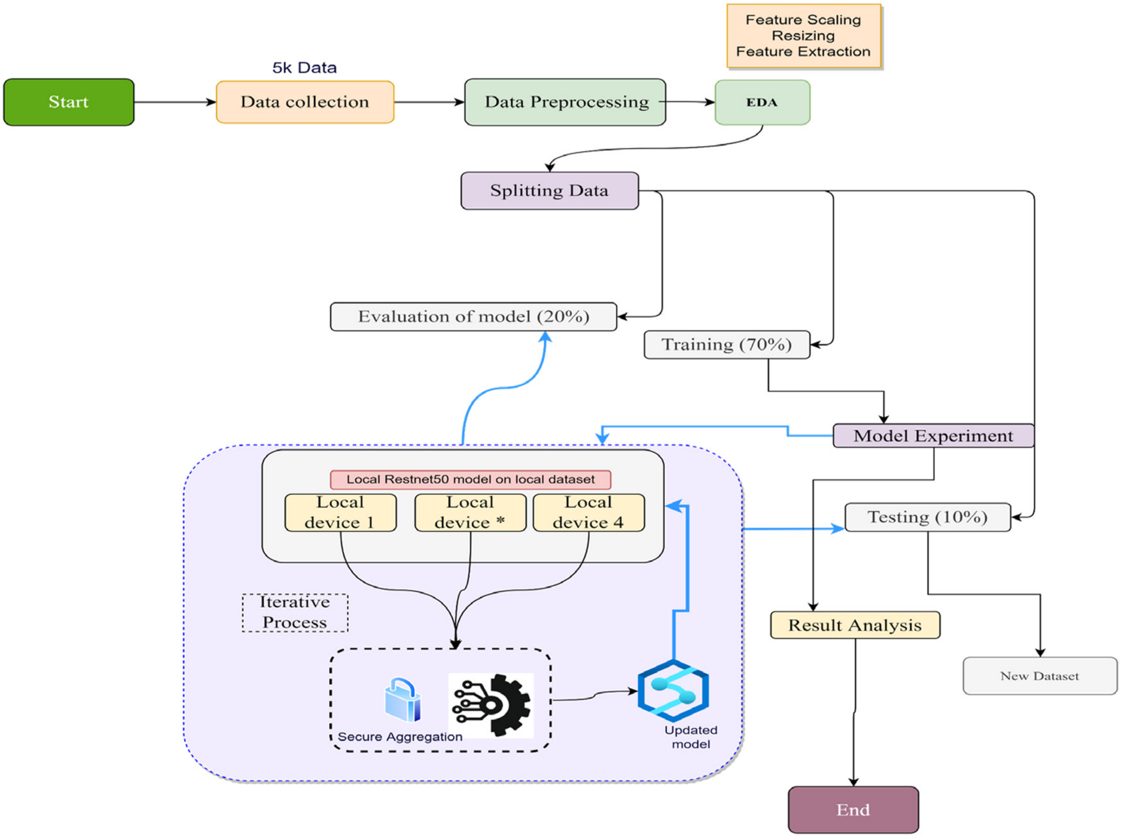

Fig. 2. The figure shows our Proposed Model. The model is followed across data splitting, training model, performing model performance analysis and using FL framework for training the model on local devices in a privacy-preserving manner.

research uses the framework of federated learning (FL) for training the machine learning (ML) models on different hospital and medical institutes. We have used state-of-the-art ML models including the convolutional neural network (CNN) based pre-trained models that include Inception, AlexNet (8-layer CNN model), DenseNet, Residual network-50 (ResNet50), and visual geometry group-19 (VGG19) for pneumonia image classification. The research is inspired to observe the machine learning model performance in a privacy preserving architecture of FL. Our research demonstrates the data privacy of contributing entities while using the data mutually following the GDPR and DPA.

2. Literature review

According to (Muhammad et al. 2021) machine learning methods can be effectively utilized for pneumonia detection [5]. In this section, we have elaborated the state-of-the-art literature done for effective medical image detection. It also covers the aspects of privacy-preserving techniques while using medical data. We have explored and reviewed the work that used machine learning techniques applied in medical image classification while keeping the data privacy. We have analysed the achieved results and elaborated on the limitations of the

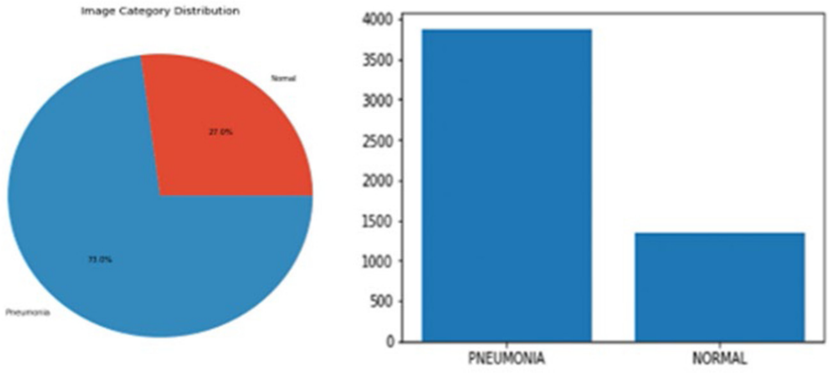

Fig. 3. Data distribution.

state-of-the-art work. Based on the research gap, we have aligned our proposed solution that can potentially fill the gap.

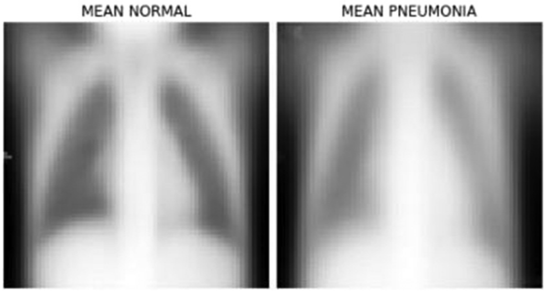

Fig. 4. Rage image.

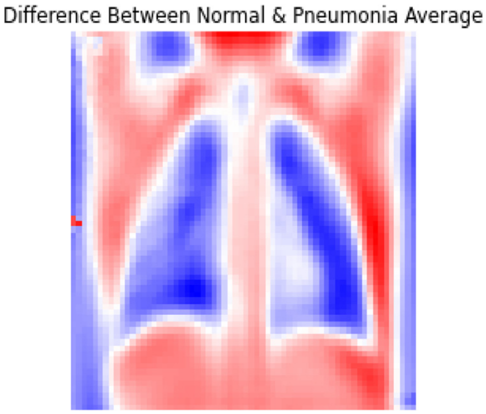

Fig. 5. Contrast between average images.

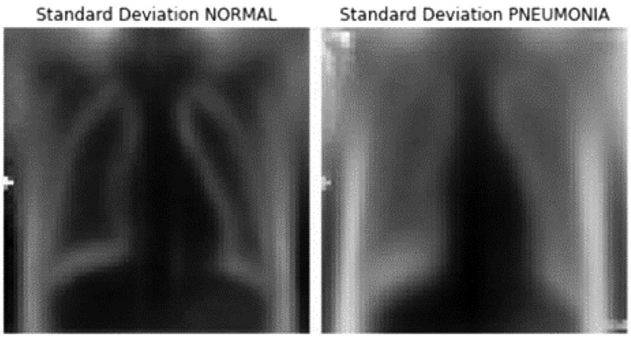

Fig. 6. Images variability.

In this chapter, we have included the state-of-the-art work done for effective medical image detection by using machine learning methods including deep learning techniques, CNN, artificial neural network (ANN), and other privacy-preserving methods.

2.1. Deep learning methods

While the problem of classification of the medical image has been quite a challenge, AI and deep learning methods made it more manageable in terms of performance and computational time. Researchers started to use deep learning algorithms on medical images as they are complex in nature. ANN is an effective deep learning method for detecting diseases such as pneumonia, tuberculosis (TB), and cancer. The researchers have performed experiment was done on the diseases using the ANN method [6]. The authors used data pre-processing techniques to clean the data and adapted the image filtration process. It achieved the sharp focus of the medical images resulting in effective medical image detection. They used a lung segmentation approach to detect features such as area, perimeters, and diameters, as well as other statistical procedures. This modification helped them to get the detection accuracy up to 92%. In this pilot study, the data was gathered from 80 individuals (patients with lung disease). The technique researchers used to reduce the size and position of the chest x-ray helped achieve the result. This experiment has produced good results in recognizing the image patterns for the classification of medical images. However, it possesses certain limitations, including changing the CXR size or position, ultimately resulting in reduced detection. The ideal method could be to use a fraction of image sizes [7]. Based on this limitation, it is essential to have a model that can keep the detection effectiveness irrespective of changes in the structure or position of the image. Researchers have widely adopted the CNN architecture for effective medical image detection which is elaborated in the next section.

2.1.1. Convolutional Neural Network (CNN)

Convolutional neural network (CNN) is one of the effective techniques for pattern recognition because CNN models are based on layering structure. The experiments [7,8] by Rasheed et al. and Sahu et al. have demonstrated that X-ray images efficiently detect and identify diseases like cancer. The authors did the study in two stages; firstly, the data cleaning was performed by removing the noises and reducing the concerned area into 65 × 65 squares. The image pixel intensities were accumulated in one file. The second stage constitutes the training, where the dataset is grouped into distinctive groups. Researchers used CNN to examine the pixels and other input feature variables. They achieved 96% while using the pixel examination method. But while using feature-based methods, they lost accuracy by 8%, which was 88%. This shows that the pixel examination method is better. However, this model ignores the feature dependency regarding real-time data. In

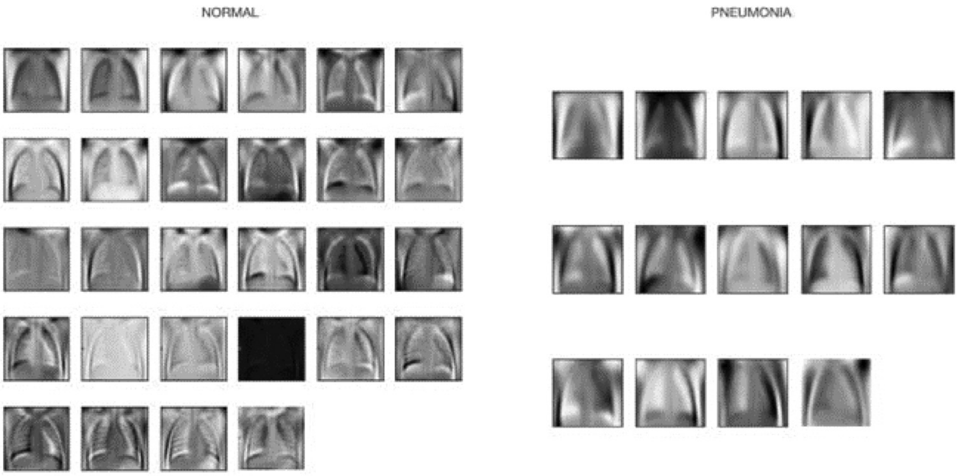

Fig. 7. Eigen image.

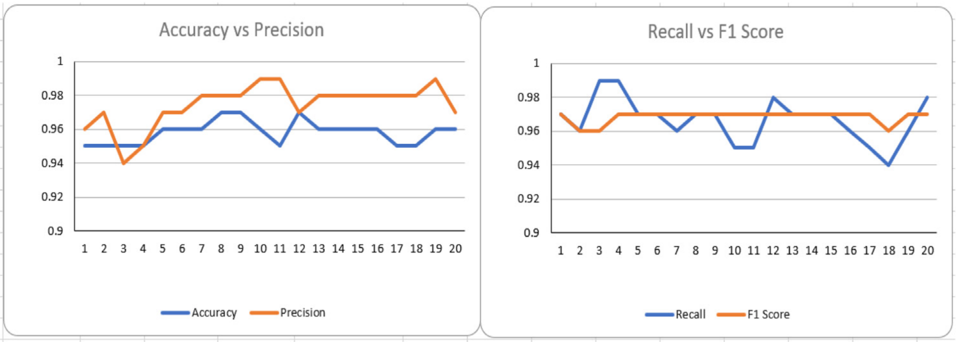

Fig. 8. Evaluation metrics using DenseNet.

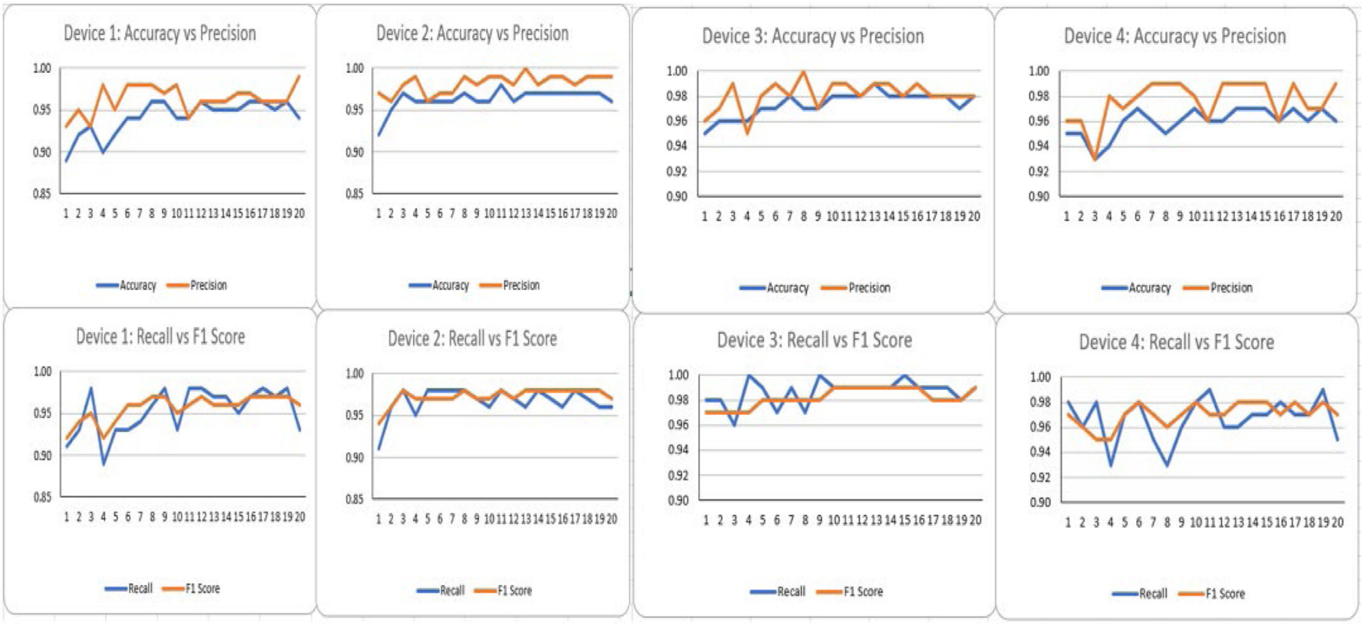

Fig. 9. Evaluation metrics using densenet with FL.

other words, the interaction with the classifier is reduced. Due to these lacking features, image ranking constitutes severe problems in terms of detection as it becomes difficult to reduce the noise and effective feature selection.

In the other experiment [9], Hira et al. used CNN to detect thorax from X-ray images. They mentioned that it is crucial to reduce the noise and enhance the area of interest for effective disease detection. They used three-branch attention gated convolutional neural network

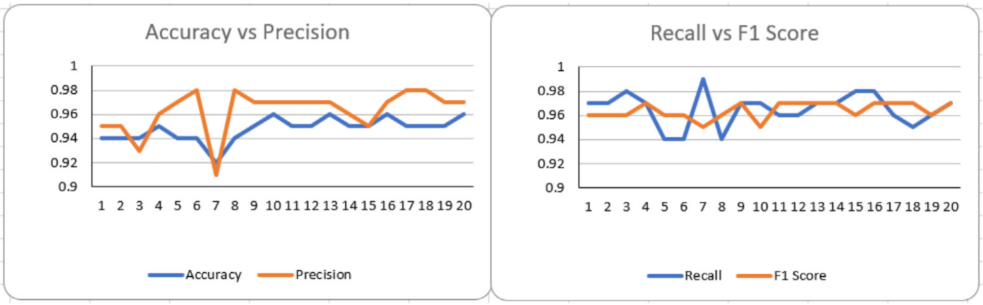

Fig. 10. Evaluation metrics using Alexnet.

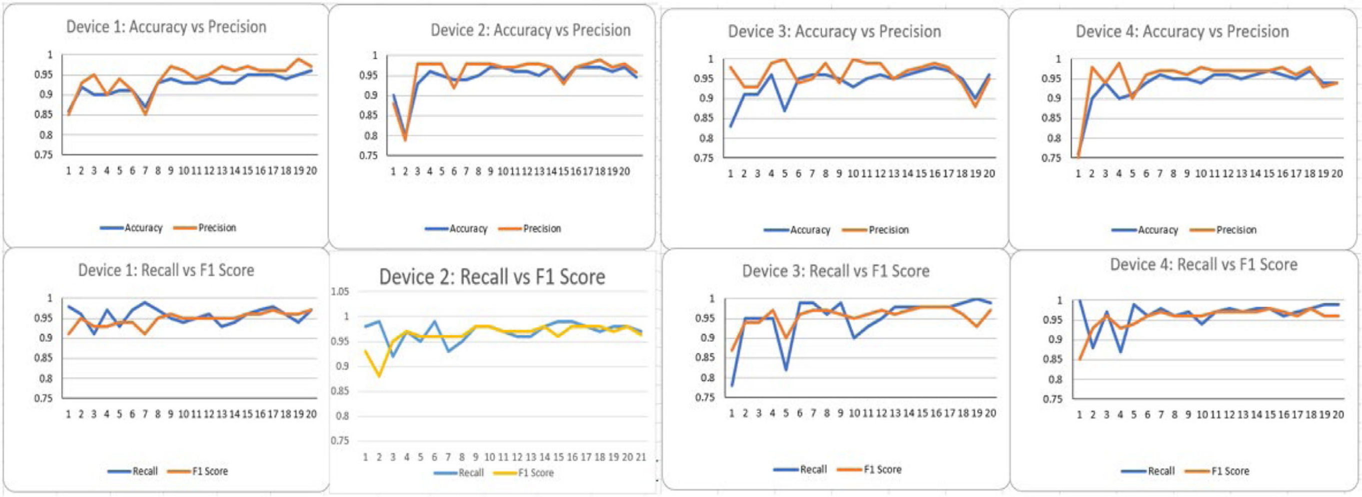

Fig. 11. Evaluation metrics using Alexnet with FL.

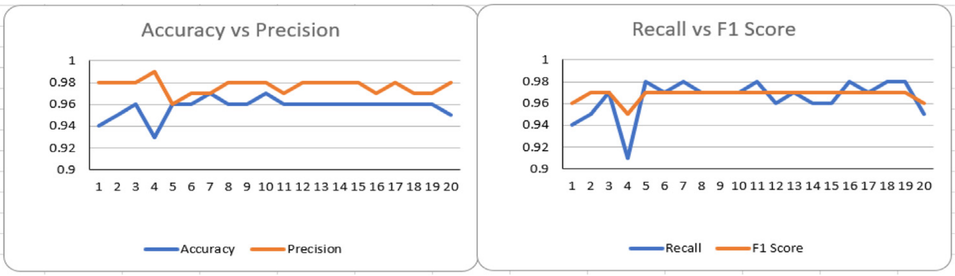

Fig. 12. Evaluation metrics using Inception.

(AGCNN) to reduce noise and effective CXR alignment. The dataset of CXR-14 was used to identify the numerous region using CNN. The researchers have used a multiple instance learning approach (MIL) that has helped to improve classification performance. The multiple stages in the research work and feature engineering have produced an effective classification result. While the experiments have achieved effective classification, the precise detection of thorax disease was ineffective. After the experiment, researchers found an area under the curve (AUC) of 0.87. It is evident that AGCNN is effective in a single stream of the dataset; however, it has reduced detection capability regarding various datasets. In other words, the changing in the parameters causes the model's inability for an effective prediction.

Another study [10] was concluded by Wahab et al. on pneumonia disease prediction using the CheXNet algorithms that constitute with 121 layers of CNN. Researchers carried out the study while collecting data from various sources. The data cleaning was performed by resizing the images into 224 × 224, and afterwards, the model was normalized and trained. The model was merged with the modified alexnet framework (MAN), which helped improve the overall model performance. The model efficiency was promising in detecting pneumonia. While the detection efficiency was promising, the image segmentation was ineffective.

Another research [11] has been done by Duong et al. on tuberculosis (TB) detection using chest X-ray images. The CNN model of AlexNet and GoogleNet was used in the experiment. The deep CNN, also known as Deep Convolutional Network (DCNN), was utilized to determine and detect the availability of any nutritious objects and other respiratory conditions. The detection of pneumonia was performed on ImageNet using the trained and un-trained model. The radiograph was used for testing and validation purposes on the specific dataset. The images obtained from the radiograph was sorted into 256 × 256 pixel. Researchers transformed it into a graphic format so it can be uploaded to the computer using the Linux operating system. Researchers achieved AUC of 0.99 in this experiment where they used radiograph images. The proposed experiment shows that the DCNN performs well in detecting

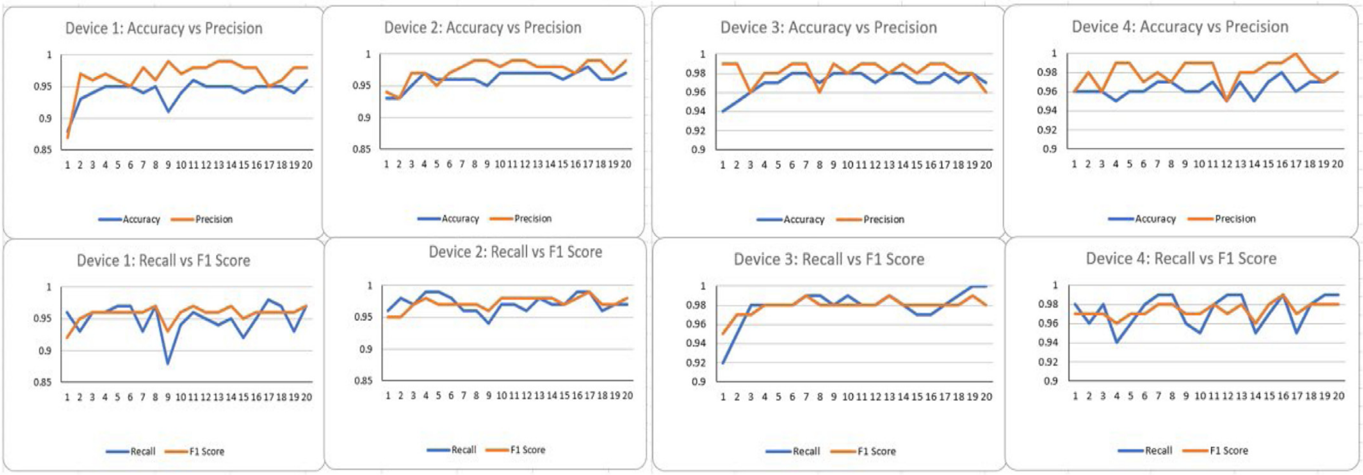

Fig. 13. Evaluation metrics using inception with FL.

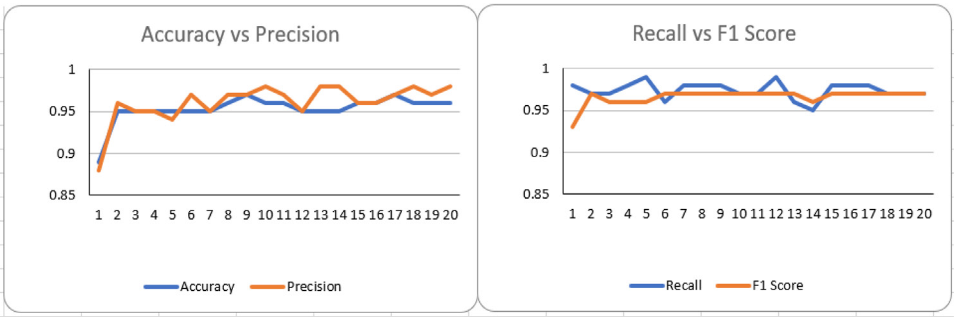

Fig. 14. Evaluation metrics using VGG19.

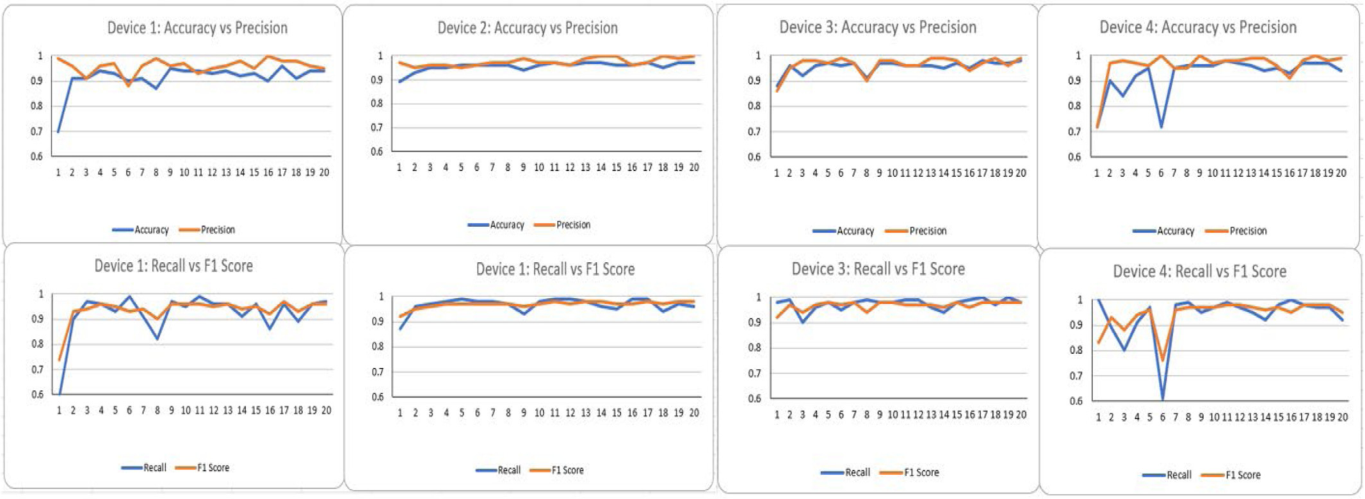

Fig. 15. Evaluation metrics using VGG19 with FL.

TB disease; however, it constitutes limitations when a larger amount of parameters are provided. Researchers did not consider that DCNN is highly processor intensive, requiring powerful machines to achieve better results.

A CNN model was proposed by Chouhan et al. to diagnose lung inflammatory disorder [12]. The authors used the dataset constituting the 14,696 images collected from 120 computerized tomography (CT) scans from multiple medical institutions. The dataset was based on multiple diseases, including TB, pneumonia, and cancer. In the experiment, researchers have used AlexNet, which is one of the CNN approaches. AlexNet is comprised of multiple layers, and in an experiment, it has achieved an accuracy of 85.5% in detection. However, the model overfitting is an issue in this research work. However, the use of AlexNet could be effective in complex dataset where it can scale up to millions of images.

In a different research proposed by Zhang et al. [13] and Wu et al. [14], a residual network (ResNet) was used to detect the two types of cancer, which are benign and malignant. The researchers achieved 92% of accuracy in cancer detection. Although the research has identified cancer in the nodule, the identification of patterns (position)

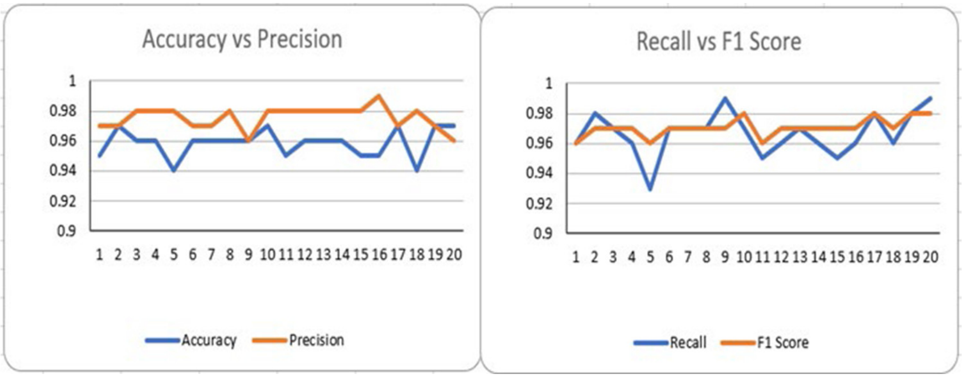

Fig. 16. Evaluation metrics using Resnet50.

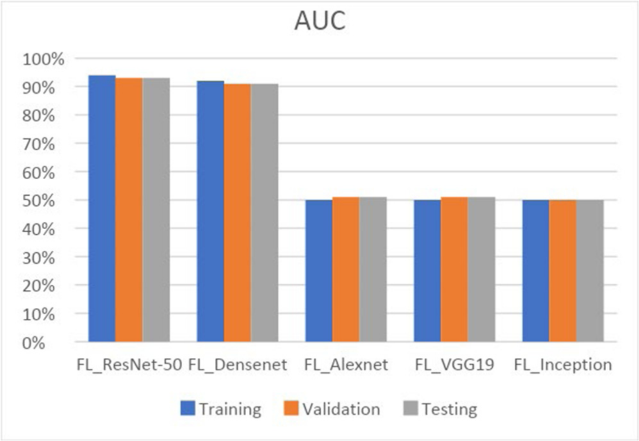

Fig. 17. AUC curve contrast among the different models.

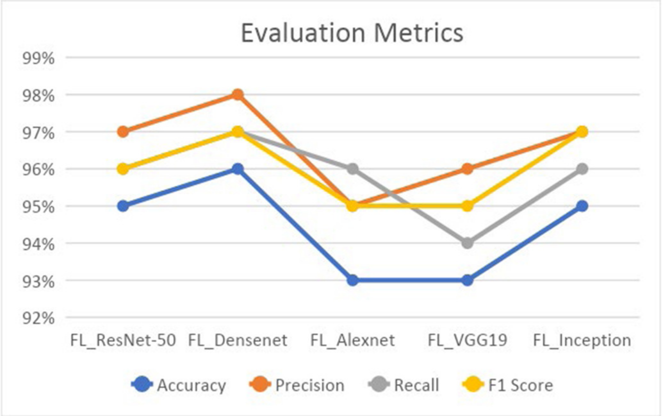

Fig. 18. Evaluation metrics of different models.

of the nodules has been the limitation of the research work. In the research, Japanese Society of Radiological technology (JSRT) dataset was considered as benchmark data for the classification of radiographs.

Wang et al. have experimented Covid-19, pneumonia, and lung opacity while using the chest X-ray scans by using the pre-trained deep learning model of Inception [15]. They achieved an accuracy of 99.72% and demonstrated the highest accuracy for X-ray disease detection while using the 21,165 chest X-ray images dataset. The researchers have used the epoch rate of 10. We conclude that ten epochs can provide better results based on this research. Our proposed experiment also used higher epochs to get better accuracy. But we found evidence that using higher epochs can overfit the model.

A comparative analysis was done by Mujahid et al. where authors performed experiments on Inception, ResNet50, and VGG-16 on the pneumonia X-ray dataset [16]. In this experiment, the Inception achieved the highest accuracy of 99.29% while using the epoch of 25 in the 7750 chest X-ray images dataset. The efficiency of the Inception model could be improved by adjusting epochs and comparative analysis. The effectiveness of the Inception model involves the flexibility of inputting multiple sizes, which is beneficial in a large variety of real-time datasets. In our methodology, we analysed the performance of the Inception model on the dataset with 20 epochs with and without a federated learning framework and performed comparative analysis.

A different approach was performed on the pneumonia image dataset (Chest X-ray) of 7,150 images using VGG19, VGG16, ResNet50, and Alexnet, along with the SoftMax classifier function. The results showed that the performance of VGG19 stands out in contrast to other models. The result also shows that the VGG19 performs better than the VGG16. The accuracy of pneumonia detection was 86.97% with VGG19, and the accuracy was enhanced by using an alternative classifier where random forest proved effective in classification with VGG19. The achieved accuracy was 97.94% as proposed by Kaissis et al. [17]. This 10.97% rise in accuracy proves that the combination of the models provides better accuracy. We also took the hybrid approach from this inspiration, which we studied in detail. We have extended our methodological research to use the VGG19 in our dataset and used it in the framework of federated learning to utilize the dataset in a privacy-preserving way. Our proposed methodology brings this research to use multiple real-time datasets with VGG19 in a privacy manner while improving the training of the model. This will help our research question to answer if we can use the machine learning models in real-time data.

The research work for multiple disease classification including thoracic disease was done by Wang et al. [15], Li et al. [18]. The achieved output of the research has indicated that the detection of thoracic disease is determined using the multi-labelled images and disease locality method. This approach is excessively adapted for thoracic disease detection. This approach helps to determine the abnormalities in the chest X-ray and also assists in locating the pathologies while using the DCNN methods. To achieve better classification, quantization methods are effective; however, the Toğaçar et al. [19] have used support vector machine (SVM) to achieve adequate detection performance. This approach involves intensive processing power and higher graphic processing unit (GPU), which is one of the drawbacks. It also comprises issues like model overfitting, which involves a higher parameter for model training. But overfitting can be overcome by using a vast amount of data.

The researcher conducted a study to differentiate paediatric chest X-ray bacterial and viral forms using CNN models. The data analysis and visualization were considered to identify the critical areas in CXR that are essential for model prediction. The analysis approach helped identify the model performance and evaluated the workflow. This research has revealed that VGG16, one of the CNN architectures, has produced effective disease detection in differentiating bacterial and viral pneumonia. While using the VGG16 approach, Alsharif et al. [20] have achieved an accuracy of 96.2%. The model has been adapted for effective performance metrics that have achieved the results' generalizations. The slower performance of the model has been the drawback of this research as the data training is time-consuming; secondly, the model occupies larger disk space and higher bandwidth.

On the other hand, the DenseNet model was used on the pneumonia image dataset to enhance classification, where researchers have compared the performance of DenseNet with CNN models. The densenet overcomes the drawback that VGG16 had, but it dropped to 92%accuracy. This accuracy was achieved in the experiment using max-pooling; however, the achieved precision and recall are not optimal. It is evident that CNN models provide better accuracy with better performance. In our methodology, we have pursued the above findings while using the optimizer, resulting in better accuracy. We have also used 20 epochs with the DenseNet model as performed in the above experiment, which has achieved better results in terms of true positive rate while doing a comparative analysis of 5, 10 and 15 epochs.

2.2. Privacy-preserving techniques

Privacy is one of the concerning areas of our research. Therefore, we have studied the different state-of-the-art literature to perform machine learning training while keeping the privacy intake for customers. Due to the GDPR law followed by data protection act 2018 [21], data sharing is impossible between institutions. This chapter will explore what other researchers have done to tackle this issue.

The research performed by Hegedűs et al. [22] shows that a novel approach for data privacy is gossip learning based on a decentralized structure. The research is done to differentiate gossip learning and federated learning (FL) performance. In the experiment, data were collected from mobile devices with good network coverage and distorted network coverage. Afterwards, the dataset was trained by the ML model individually on gossip learning and FL. While considering data privacy, the parameters of scalability, semi-centralized and instantaneous behaviour, FL has performed effectively in contrast to gossip learning. The overall performance of gossip learning was reduced while data modelling, the confined size of the data, and scalability were also negatively impacted. It proves that federated learning could result in better performance to achieve an efficient system for medical image detection while using real-time datasets; it is essential to have an effective method for data privacy that has higher scalability and is instant in operation.

The applications of FL are quite limited because it was recently introduced. An experiment was performed by Lee et al. [23], where the FL is applied to the electronic health record (EHR) to predict diseases. The research is followed by training the data on an individual hospital and medical institution. Effective results were achieved by training the model locally. In this approach, data was not shared. Instead, the trained model from individual entities is aggregated to form a centralized model. As this is one of the efficient ways of training the model from the real-time data while preventing data share, thus promising privacy, we are considering this approach in our research. This approach is scalable in the way that the repetitive cyclic process helps the model to learn the new patterns from the data that fix the problem of data heterogeneity.

To tackle privacy, researchers worked on blockchain technology as well. It is another technology for ensuring data confidentiality. It is based on a decentralized approach while following the cryptographic algorithm. Research has been conducted using blockchain technology in medical and transaction data [24]. This experiment achieved an effective outcome by keeping the data privacy where the cryptographic methods were deployed based on blockchain technology. Although the research is promising while considering data privacy, however, the process is slower in operation as well as processor intensive. It also causes scalability problems that are ultimately ineffective for deploying in real-time datasets as highlighted [25]. Therefore, considering blockchain technology for real-time medical image data in hospitals and medical institutes is not a prominent solution for training the model.

A privacy-preserving approach for detecting Covid-19 from X-ray images was proposed using real-time data from medical research canters. In the experiment, researchers used blockchain technology for data authentication and a federated learning approach to training the model globally. The model was based on deep learning techniques. This research used the capsule network approach for covid-19 detection and normalization techniques as the data were heterogeneous (sourced from various research communities). The experiments performed by Kumar et al. [26] have produced higher accuracy of 98.68% in detecting covid-19 compared to other state-of-the-art. Although the experiment has produced higher accuracy, however, the process is slower, caused of the blockchain ledger. On top of that, scalability is also an issue. The sole consideration of federated learning in the model can produce promising results regarding data privacy as it will ensure faster processing and broader scalability.

2.2.1. Differential Privacy (DP)

The concept of differential privacy (DP) for medical image analysis was proposed by Ziller et al. [27]. This algorithm finds the overall patterns recognized instead of knowing the particular feature. The authors have used the algorithm of DP and stochastic gradient descent (SGD), termed DP-SGD. The use of SGD algorithms has produced higher privacy when compared with the existing state-of-the-art. The authors used the paediatric pneumonia datasets framework for image classification and segmentations. Researchers concluded that using DP in the neural network can possibly be done while having reasonable classification and segmentation performances. The work has given a direction to consider the concepts of DP for data privacy guarantee while doing medical image analysis. The use of SGD has lacking in handling larger datasets. Our method involves multiple hospitals and medical canters collaborating mutually; therefore, the dataset size increases. The SGD process involves continuous updates of the training datasets, which is computationally intensive; therefore, it is not feasible for larger datasets.

2.2.2. Federated Learning (FL)

Federated learning (FL) is a framework introduced by Google that allows machine learning models to get trained on the dataset in a privacy preserving manner. In this approach, there is centralized server that is used as source and destination for ML models. All the participants are connected to the central server in the way that server send the model to individual participants for training the data. Once the model is trained at local participants, the trained model is aggregated at central server where the data is not shared. There have been multiple research projects carried out to use this framework in different domains.

A study was conducted by Pfitzner et al. [28], where comparative analysis was performed in FL. The researchers have described the potential of FL in the medical sector, where patient data privacy is a big challenge. The comparative analysis shows that GDPR is harder to maintain when it comes to data sharing for research purposes. The issue has been resolved by the capabilities of FL, where models can be used in medical data that can potentially ensure data privacy. The authors have highlighted some of the challenges involved, including hyperparameter optimization and encryption in the FL framework. The identified hyperparameter challenge can be solved using deep neural network techniques.

In contrast, the encryption challenge can be solved using a secure aggregation protocol that allows users to perform encryption in their data. Our proposed work will use transfer learning mechanisms that effectively optimize hyperparameters and encryption. The learning process of the model is enhanced in the broader spectrum once FL is used along with the transfer learning.

The framework of FL was proposed by Feki et al. [29] for the detection of covid-19 and normal X-ray images using the client–server approach. The authors have used VGG16 and ResNet50 for the feature extraction and the classification of images. The model takes the images and displays the output as the probability of covid-19 infection. The results achieved show the same performance of both models in detecting the probability of covid-19. This shows that VGG16 and resnet50 in medical image detection can improve performance. Collaboratively using pre-trained CNN models could be breaching the data privacy as there is no encryption on the ML algorithm; however, FL has a secure aggregation protocol that could save patients' privacy. Using the FL framework helps to achieve the adequate performance of the model training while following the cyclic process between client and server.

The experiments were performed by Zhang et al. [30] on the CT scan images to detect Covid-19. In the experiment, datasets have been used from multiple sources using the approach of FL, where the global model is trained with respect to the combined approach of local models from individual clients. To ensure the effective communication and performance of the model, the researchers have used the dynamic fusion method to enhance the FL model. The results demonstrated that the dynamic fusion approach performs well with respect to the original setting of FL while considering privacy, stability, and communication. Although the dynamic fusion technique performs well for increasing the efficiency of FL, using the transfer learning algorithms, for instance, GhostNet used in the experiment, takes longer to train the model. In the real-time implementation, there should be an approach that speeds up the time process for optimal outcomes.

Kaissis et al. [17] have introduced an approach of using encryption for training real-time medical image data by the use of FL. They have used deep learning models for training the data. In the proposed work, the model is sent across to local clients; however, the model gets encrypted along with data while training. This way, only the encrypted trained model is shared with the central server. The results show that this concept can be applied to multiple datasets using the privacy-preserving approach. The encryption model and federated learning can ultimately lead to securely using the model over the public Internet. Although the encryption method proposed by the researchers gives promising results for encrypting the training model over the public internet, however, the process of encryption and decryption is processor intensive as well as time-consuming. Federated learning already has an inbuilt encryption method known as secure aggregation protocol as elaborated by Bonawitz et al. [31]. A secure federated learning ensures data encryption at all levels of its process.

The concept of FL has been applied to many areas of medicine. Zhang et al. have worked on using FL in biomedical monitoring data [32]. The researchers used heart data in this experiment with smart bands on the wrist. The experiments have produced promising results while considering the data privacy of the participants. The concept was used in the context of IoT, where the data collection was done with the internet of medical things (IoMT). The results have produced better performance accuracy in contrast to the traditional model. The process was performed by using deep learning algorithms. The idea of FL has brought a massive revolution in the medical system, where patient data privacy is a serious consideration. The researchers have used the concept of FL in biomedical monitoring data that has produced the result of 87.55% accuracy while combining the two datasets (1 and 2) while individually, they have produced the individual accuracy of 81.65% and 84.08%, respectively. It shows that the performance of FL increases as the datasets are combined. We are using a similar concept of combining multiple data from hospitals and medical institutions to achieve effective medical image detection results.

A study has been conducted to use the framework of FL in the cancer research [33]. The study has explored that there are quite few numbers of research that is conducted on the cancer diagnoses. It has well demonstrated the use of using the FL architecture in the prediction of disease like cancer. It has also illustrated that the 56% of the research is conducted using the cancer datasets while the rest has been used for benchmarking.

In terms of the performance of FL, researchers have performed different types of experiments to demonstrate the effectiveness of this architecture. In research [34], the experiments were conducted to detect the cardiac arrhythmias where the achieved accuracy was 98.9%, however another research was conducted by using the similar approach of FL and dataset with distinct models that has achieved the accuracy of 87.85% [35]. Similarly, if we analyse the results [36,37], the area under the curve (AUC) has come up as 0.78 and 0.89 respectively. These results achieved are not as effective in using them in cardiovascular disease (CVD) dataset, although the results achieved in [34] are comparatively better, however for the other experiments as mentioned above, the research conducted on the CVD dataset has achieved the accuracy of 91% which is 12 months prior to the disease [38].

In contrast, if we analyse the performance of FL in the diabetes diagnosis, the research conducted has achieved the accuracy of 99.24% which is slightly higher than the relative models used in the field [39]. Furthermore, in the review study [40,41], authors have demonstrated that the results achieved through the traditional ML models can be comparable to the deep leaning models. In contrast the results achieved in the research [42], are not as higher in comparison with the achieved accuracy of 72% that is lower than the traditional ML models. The researchers have demonstrated that the use of deep learning models enhance the performance of FL [43], while in contrast, another research [44] has demonstrated that the FL has better generalization in comparison than the models itself. In summary, FL could outperform while using the ML models, however the results achieved are not always the thumb of rule that ensure the hypothesis of correct disease prediction. Although, the architecture of FL quite accurate and achievable, however in some cases the ML models outperform in term of disease prediction accuracy even if the privacy is ignored.

Based on the above literature review, we can understand that the framework of federated learning is effective in using the real-time datasets in a privacy-preserving manner. The comparative analysis of different state of artwork in image classification have shown that CNN-based pre-trained models are effective in medical image classification. We are taking this motivation forward and experimenting the pre-trained models of AlexNet, DenseNet, ResNet-50, Inception and VGG19 in federated learning framework to observe the performance of pneumonia image detection. Next chapter illustrates our research methodology that is based on the FL framework.

3. Proposed methodology

We have used the framework of federated learning in our research work as it reflects the training of the machine learning models at different medical institutions and hospitals while keeping the data privacy. In our proposed framework, we have used 4 different virtual devices corresponding to each entity where the ML model will get trained separately. In this approach the training data is split across equally and models gets trained. Our research model design is demonstrated as below:

The above research model design follows the data collection process that includes both pneumonia and non-pneumonia images. Afterwards, the data pre-processing ensures the images standard approach before modelling and also the data goes through the stages of exploratory data analysis (EDA) to get insights of the data. Once, the data pre-processing is completed, the data is split into training, validation as well as testing in the ratio of 70%, 20% and 10% respectively. The training data is used to train the ML model where the model hyperparameter can be adjusted as well as the optimizer is added. The training data is sent across the equally to the virtual devices as shown the above figure.

Now, the untrained machine learning model is sent to each participant and it gets trained locally on devices. Once the model is trained, it is sent back to the federated learning aggregated server where all other models from the individual devices are combined together. In this approach, data is not sent across, only the trained model is received. Once all trained models are combined together, the central aggregated model is sent back to each device and process is repeated. In this way, the model learns the maximum feature variables of the data. The model transmission is controlled by the protocol used by FL knows as secure aggregation that encrypts the model and avoid reverse engineering. After the allocated iterations the model is validated by using the validation dataset, if results need to be improved, the process is repeated with adjusting ML model hyperparameters. Once the validation test is completed, the model is tested by using the test dataset.

3.1. Data collection and pre-processing

In the field of medical image detection, deep learning algorithms have been widely used. To ensure the efficiency of the algorithms, chest X-rays (CXR) have been effective solution for researchers to perform the experiments. In our research work, we have used CXR dataset of lungs having pneumonia/non-pneumonia images. The dataset is open sourced provided by the University of California San Diego [45]. The available dataset is cleaned by the professionals to keep only the high-quality images. We have selected the above dataset due to the fact that, it belongs to the authentic source which is reliable and accurate. Also, the results achieved while using this dataset can be replicable and comparable with the other research on the similar dataset. There is number of benchmark available for this dataset on Kaggle and several research have also been done while using this dataset. The choice of this dataset reflects the relevancy to our research (containing binary classes), high quality, publicly available as well as widely used by other researchers and well recognized in the research community.

The dataset itself constitutes of 5856 CXR images in JPEG format. In terms of images classification in the dataset, 27% is classed as the normal (non-pneumonia) and rest 73% is classified as the pneumonia. Due to the higher amount of pneumonia images, the data augmentation process is carried out that involves rotation, zoom and shifting of widths and heights.

Once the data collection is done, the data pre-processing is the initial stage remove any irregularities. We have resized all images into 224 × 224 pixels. Due to data imbalance between the pneumonia and non-pneumonia images, we have performed data augmentation. We have used the dataset that is labelled and processed it by image segmentation.

3.2. Exploratory Data Analysis (EDA)

EDA is an essential stage in the data analysis as it helps to understand the insights of the data. It differentiates the important elements inside the dataset, so those can be targeted to achieve effective results. EDA has helped to analyse the underlying patterns in the data and distinguish the trend and patterns between them. In our CXR dataset, EDA gave us insights of the data which we have used for the selection of effective ML model. The EDA of our research constitutes of below stages:

3.2.1. Data distribution

Our CXR dataset contains two divisions, pneumonia and non-pneumonia. The classification can be represented in the below graphs:

As it demonstrate the above pie chart, 27% of the dataset contains non-pneumonia and 73% as pneumonia. This stage of EDA gives us clear understanding of the data imbalance where the pneumonia images are far more than the non-pneumonia images.

3.2.2. Rage image

This phase of EDA has given us the average image based on each class. Rage basically calculates the average pixels of all available pixels of the give class. Therefore, in our case of CXR, the rage image gave us average value of available pixels on each class as it can be seen below:

This EDA gives us the idea that average pneumonia images shows higher obstructions in the chest side unlike normal images that shows higher clarity.

3.2.3. Contrast between average images

While considering the average image view, the contrast between the images that includes pneumonia and normal images shows the illustration that distinguish the common and different parts. The contrast between the normal and pneumonia image of our dataset is illustrated below:

As it can be observed from the above figure that there is more common area on the left lung (higher blue shade) in contrast to the right one.

3.2.4. Variability

Variability is also one of the effective EDA tools to distinguish between the classes. It is used to calculate the variance or standard deviation between the classes (pneumonia or normal). The variability of our dataset is demonstrated as below:

The above figure demonstrates that, there is higher variability in the pneumonia images in contrast to the normal ones.

3.2.5. Eigen images

We have used the dimension reduction techniques for better visualization of the data. To differentiate between the classes, we have used principal component analysis (PCA). PCA has helped to achieve the better insights of the image. Eigen image is used in PCA that is involved in image reshaping into metrics that can ultimately used for better visualization. Below figure demonstrates the PCA that constitutes the 70% of the variability based on single class of the dataset:

The above figure shows that the eigen image of the normal image has higher definition in terms of the image structure.

We have elaborated tools and techniques used for data analysis, while keeping this in place, in the next section, we will demonstrate our experiments and results while using CNN based pre-trained models individually as well as with federated learning.

4. Models and experiments

In our research, we have conducted experiments while using the CNN based pre-trained models. We have used deep learning models of Resnet-50, AlexNet, Densenet, Inception and VGG19. We have selected these models based on the analysis from the literature review that can possible fit best in the federated learning architecture for the medical image detection. CNN based models are highly scalable and effective in training. It also prevents the overfitting issues due to bigger dataset.

DenseNet architecture constitute of layering blocks where the layers are interconnected. ResNet50 on the other hand constitute of 50 layers that involves skipping connections and shortcuts which is widely adopted in medical image classification. VGG19 involves multiple layers of CNN that are fully interconnected and hence increases the layering depth that ultimately helps in better image detection. Alexnet involves eight layers and highly effective in complex datasets that are highly processor intensive. Inception is one of CNN-based pre-trained model that is takes less computation power and avoids overfitting. We have considered these models due to their individual efficiency and literature analysis that can be used effectively in medical image detection.

In our experiments, we have used the 4 virtual clients where the dataset is equally distributed. We have selected 4 virtual devices as a round number based on the analysis from the previous work. The

Table 1
Evaluation metrics using densenet with federated learning.

	Accuracy(%)	Precision (%)	Recall (%)	F1 Score (%)
Device 1	94%	96%	95%	96%
Device 2	96%	98%	97%	97%
Device 3	97%	98%	99%	98%
Device 4	96%	98%	97%	97%

Table 2
Comparative analysis using densenet with FL.

Models	Training	Validation	Testing
DenseNet	99%	96%	95%
FL_DenseNet	92%	91%	91%

Table 3
Evaluation metrics using Alexnet with FL on individual devices.

	Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)
Device 1	92%	94%	95%	94%
Device 2	94%	95%	97%	96%
Device 3	93%	96%	95%	95%
Device 4	93%	95%	96%	95%

research can be extended to add more number of devices to observe the performance of the ML models in federated learning framework. In terms of CNN based pre-trained models, we have set the epoch value to 20. Epoch is basically the forward and backward propagation in neural network. Higher number of epoch can result in model overfitting and produce ineffective results. Similarly, lower value results in underfitting the model. Therefore, as per the analyse from the literature, 20 has been an effective in most of the cases for medical image classification. We have conducted 10 different experiments while using our dataset in the following sequence:

- DenseNet and DenseNet with FL (FL_DenseNet)
- AlexNet and AlexNet with FL (FL_AlexNet)
- Inception and Inception with FL (FL_Inception)
- VGG19 and VGG19 with FL (FL_VGG19_FL)
- VGG19 with FL (FL_VGG19_FL)
- ResNet-50 and ResNet with FL (FL_ResNet50)

In the next section, we have analysed our results based on the experiments.

4.1. DenseNet Vs FL_DenseNet

DenseNet

We have used the DenseNet model on the dataset with the set epoch of 20 and we have achieved the mean accuracy, precision, F1-Score and recall of 95%, 97%, 96% and 6% respectively. The results can be observed in the below graph:

FL_DenseNet

In FL architecture, we have used four virtual clients named as Device 1, Device 2, Device 3 and Device 4. The trend of 20 epoch has been followed as well on the individual device. The evaluation metrics can be observed in the below table (see Tables 1–12).

The trend of evaluation metrics can be analysed in the following graphs:

Discussion (DenseNet Vs FL_DenseNet)

We have compared the performance of DenseNet model with the FL_DenseNet. Following table shows the comparative analysis of training, validation and testing dataset by using the evaluation metrics of area under the curve (AUC) that demonstrate the higher true positive rate.

The above table shows that the value of AUC is higher in simple DenseNet model than the FL_DenseNet. However, there is less difference among the individual virtual device. Therefore, it can be understood that in federated learning framework where more than one device devices are used, the value of AUC is slightly dropped.

Table 4
Comparative analysis using Alexnet with FL.

Models	Training	Validation	Testing
Alexnet	97%	95%	97%
FL_Alexnet	50%	51%	51%

Table 5
Evaluation metrics using inception with FL based on individual devices.

	Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)
Device 1	94%	96%	94%	95%
Device 2	96%	97%	97%	97%
Device 3	97%	98%	97%	97%
Device 4	96%	97%	97%	97%

Table 6
Comparative analysis using inception with FL.

Models	Training	Validation	Testing
Inception	98%	97%	96%
FL_Inception	50%	50%	50%

4.2. AlexNet Vs FL_AlexNet

AlexNet

We have performed experiments over the e20 epochs while using the AlexNet model on the provided pneumonia dataset. We have achieved the mean accuracy, precision, recall and F1 score of 94%, 96%, 96% and 96% respectively. The following trend shows the evaluation metrics results throughout the 20 epochs:

FL_AlexNet

While using the AlexNet model on the FL framework, we have achieved the following results over the four device and mean value of 20 epochs:

The trend of individual epoch can be observed on the below line graph over the four devices:

Discussion (AlexNet vs FL_AlexNet)

According to the experiment results, the individual training, validation and testing dataset has almost same results either in AlexNet or FL_AlexNet, however when we observe the evaluation metrics results between the AlexNet and aggregated FL_AlexNet, the results have high differences. The values of AUC can be observed in the below table:

The results in comparison shows that the use of FL_AlexNet is not been an effective approach for pneumonia image classification as due to high values of false positive and false negative. However, for the simple AlexNet, the classification has been observed effectively.

4.3. Inception vs FL_Inception

Inception

Inception is one of the CNN based pre-trained model used widely in image classification. We have used the inception model on the dataset for training. The mean average of 20 epochs has achieved the accuracy, precision, recall and F1 score of 95%, 97%, 96% and 96% respectively. The individual trend over the 20 epochs can be observed in the following graph:

FL_Inception

We have used the inception model over the FL framework over the 20 epochs. The mean evaluation metrics can be observed in the below table:

Table 7
Evaluation metrics using VGG19 with FL.

	Accuracy(%)	Precision (%)	Recall (%)	F1 Score (%)
Device 1	91%	95%	92%	93%
Device 2	95%	97%	96%	96%
Device 3	95%	96%	97%	96%
Device 4	92%	96%	93%	94%

Table 8
Comparative analysis using VGG19 and with FL.

Models	Training	Validation	Testing
VGG19	99%	97%	95%
FL_VGG19	50%	51%	51%

Table 9
Evaluation metrics using resnet50 with FL based on individual devices.

	Accuracy(%)	Precision (%)	Recall (%)	F1 Score (%)
Device 1	93%	96%	95%	95%
Device 2	96%	98%	96%	97%
Device 3	97%	97%	98%	97%
Device 4	96%	98%	96%	97%

Table 10
Comparative analysis using resnet50 and with FL.

Models	Training	Validation	Testing
ResNet-50	99%	95%	94%
FL_ResNet-50	94%	93%	93%

The individual trend of 20 epochs can be observed in the below graphs over the four devices:

Discussion (Inception vs FL_Inception)

As seen on the above figures, the evaluation metrics on the individual devices over the FL framework is higher than the simple inception model. However, while comparing the inception with the FL_Inception, the below tables shows the AUC results on training, validation and testing dataset:

The above results shows that, although the value of evaluation metrics is higher individually on the devices, the aggregated model is ineffective as the results are half the value of simple inception. In other words, there are higher false positive and false negative rates, therefore the use of FL_Inception is ineffective in classifying disease.

4.4. VGG19 Vs FL_VGG19

VGG19

We have used VGG19 model to train our dataset. In the experimental results, we have achieved the mean accuracy, precision, recall and F1 score of 95%, 96%, 97% and 96% respectively throughout the 20 epochs. The trend on the individual epoch can be observed in the below line graph:

FL_VGG19

We have trained the VGG19 model in a FL framework over the four devices. The mean accuracy, precision, recall and F1 score can be observed in the below table:

The trend on the individual epoch can be observed in the below chart over the 4 devices:

Discussion (VGG19 vs FL_VGG19)

In case of simple VGG19 training of the data as well as the training on the individual device on the FL_VGG19, the evaluation metrics results seem to be similar with few exceptions. However, the following trend is observed while considering the AUC over the training, validation, and testing dataset:

Table 11
Average result contrast between different data splits.

Models	Training	Validation	Testing
FL_ResNet-50	94%	93%	93%
FL_DenseNet	92%	91%	91%
FL_Alexnet	50%	51%	51%
FL_VGG19	50%	51%	51%
FL_Inception	50%	50%	50%

Table 12
Different evaluation metrics of combined models.

Models	Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)
FL_ResNet-50	95%	97%	96%	96%
FL_DenseNet	96%	98%	97%	97%
FL_Alexnet	93%	95%	96%	95%
FL_VGG19	93%	96%	94%	95%
FL_Inception	95%	97%	96%	97%

It shows that the performance of aggregated FL_VGG19 decreases in contrast to training the dataset on simple VGG19. Therefore, it can be concluded from the results that the FL_VGG19 is ineffective in terms of disease classification as the devices in the FL framework is increased.

4.5. ResNet-50 vs FL_ResNet-50

RestNet-50

In our experiment, we have used the ResNet-50 for training the dataset. We have achieved the accuracy, precision, recall and F-1 score of 95%, 97%, 96% and 97% respectively. We have set the epoch of 20 like other models. The trend of evaluation metrics can be observed in the below graph:

FL_ResNet-50

We have trained the resnet-50 model in a federated learning framework over the four devices. The result of evaluation metrics can be seen in the below table:

Discussion (ResNet-50 vs FL_ResNet-50)

In the experiment results, it has been observed that some of the devices shows better evaluation metrics in FL_ResNet-50 contrast to ResNet-50. The comparison based on the mean AUC can be observed in the below table:

Although the above result shows that the AUC in FL_ResNet-50 is lower by 1% in validation and testing dataset, however this difference can be negligible on more complex dataset. The results achieved by using the FL_ResNet-50 is quite effective and close to the results achieved by the ResNet-50.

4.6. Evaluation of model selection

While comparing the above models in terms of performance in a FL framework, it can be observed that the FL_ResNet-50 has achieved the better evaluation metrics in contrast to the FL_DenseNet, FL_AlexNet, FL_VGG-19 and FL_Inception. The AUC comparative analysis can be observed in the below table in a descending order from top to bottom:

Following can be demonstrated visually in a bar graph:

The trend of FL_ResNet-50 stands out in contrast to the other models. The evaluation metrics can be observed in the below line graph:

It can be represented in the table as below:

In the above illustration of the evaluation metrics, FL_ResNet-50 and FL_DenseNet shows almost similar performance. The performance of the model individually on the dataset brings good results in classification, however in the FL framework, the performance has been decreased. FL_ResNet-50 has achieved highest true positive rate with the FL_DenseNet with the slightly less in contrast.

4.7. Significant test

While considering the higher true positive rate of FL_ResNet-50 and DenseNet, we have performed significant test to understand the difference in the evaluation between these two models really makes major difference. We have calculated the mean and standard deviation of the accuracy score of each model in federated learning framework. We have used T-test to determine if there is a significant difference in the accuracy scores and Wilcoxon signed-rank test.

Accuracy score reflects the total true positive and true negative out of the total predictions, while wilcoxon signed-rank test suggested the significant difference between resnet-50 and densenet in terms of statistical distribution.

The results of the analysis are as follows:

The mean accuracy for DenseNet is 0.9628 with a standard deviation of 0.0089, while the mean accuracy for ResNet is 0.9696 with a standard deviation of 0.0049.

Using a t-test, the t-statistic is −1.9974 and the p-value is 0.1097. Alternatively, using a Wilcoxon signed-rank test, the Z-statistic is 9.0000 and the p-value is 0.1097.

Since the p-values from both the t-test and the Wilcoxon signed-rank test are greater than the significance level of 0.05, there is no significant difference between the accuracy scores of the DenseNet and ResNet-50 models. Therefore, we cannot claim that one model is significantly better than the other based on their accuracy scores.

Thus, based on the accuracy and Wilcoxon signed rank test, the results achieved using resnet-50 and densenet are negligible and does not make significant difference in terms of performance. It shows that using densenet and resnet-50 can make similar performance in training the data in a federated learning environment.

5. Conclusion

Our research work demonstrates the use of ML models for effective classification of pneumonia images. The comparative analysis showed that the ResNet-50 and DenseNet model performed better in contrast to VGG19, Inception and AlexNet while considering the confusion matrix of accuracy, recall, precision and F1 score. And, also the performance is justified by using the AUC curve and compared the true positive rate among the different models and explained in the models and experiments section. We have demonstrated the use of FL framework for data privacy that allows multiple entities to share the data for machine learning model training. Our experiment results shows that the mutual collaboration of hospital and medical institutes is possible in FL framework. The research demonstrates the use of real-time dataset in a privacy-preserving manner with the highest possible accuracy. Our research can be further explored by using the different disease datasets. Although, FL provides an architecture for data security, however when it comes to its implementation in real-time in healthcare, it can encounter challenges. Below are some of the limitations and challenges of our proposed methodology for health care data security:

- In the shared model among the different hospitals and medical institutions, the data quality plays a major role. If the data is corrupted or poor in one or more parties, the performance of the ML models can be negatively impacted.
- The communication overhead is also one of the major challenges in implementing the FL framework in real-time as distraction in communication between the collaborating bodies having larger dataset can impact the model performance.
- The healthcare regulations are always in the pace of evolution; therefore it can be challenging to maintain the architecture in accordance with regulations.
- There is also the lack of standardization when it comes to the implementation of FL framework in real world.
- Currently, there is no regulatory compliance of the proposed methodology in the healthcare industry, therefore it would require additional compliance strategies which can be complex and expensive.

While having the challenges of implementing the proposed methodology in real-time, the further evolution of research can bring up the great changes in the medical field where the traditional approach of disease detection could be replaced by more ML techniques where data from other entities would be used while preserving GDPR and DPA that can ultimately assist in early disease detection and saves maximum lives.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Link is shared on article.

Acknowledgments

We acknowledge the University of Bedfordshire for providing funding support for this research publication.

References

[1] O. Ruuskanen, E. Lahti, L.C. Jennings, D.R. Murdoch, Viral pneumonia, Lancet 377 (9773) (2011) 1264–1275, http://dx.doi.org/10.1016/S0140-6736(10)61459-6.

[2] Pneumonia — no child should die from a disease we can prevent, Our World in Data. https://ourworldindata.org/child-deaths-from-pneumonia.

[3] S.Z.H. Naqvi, M.A. Choudhry, An automated system for classification of chronic obstructive pulmonary disease and pneumonia patients using lung sound analysis, Sensors 20 (22) (2020) http://dx.doi.org/10.3390/s20226512.

[4] L. Zaffiri, J. Gardner, L.H. Toledo-Pereyra, History of antibiotics. From salvarsan to cephalosporins, J. Invest. Surg. 25 (2) (2012) 67–77, http://dx.doi.org/10.3109/08941939.2012.664099.

[5] Y. Muhammad, M.D. Alshehri, W.M. Alenazy, T. Vinh Hoang, R. Alturki, Identification of pneumonia disease applying an intelligent computational framework based on deep learning and machine learning techniques, Mob. Inf. Syst. 2021 (2021) e9989237, http://dx.doi.org/10.1155/2021/9989237.

[6] Tilve S. Nayak, S. Vernekar, D. Turi, P.R. Shetgaonkar, S. Aswale, Pneumonia detection using deep learning approaches, in: 2020 International Conference on Emerging Trends in Information Technology and Engineering (ic-ETITE), 2020, pp. 1–8, http://dx.doi.org/10.1109/ic-ETITE47903.2020.152.

[7] M. Sahu, R. Dash, A survey on deep learning: Convolution neural network (CNN), in: Intelligent and Cloud Computing, Singapore, 2021, pp. 317–325, http://dx.doi.org/10.1007/978-981-15-6202-0_32.

[8] J. Rasheed, A.A. Hameed, C. Djeddi, A. Jamil, F. Al-Turjman, A machine learning-based framework for diagnosis of COVID-19 from chest X-ray images, Interdiscip. Sci. Comput. Life Sci. 13 (1) (2021) 103–117, http://dx.doi.org/10.1007/s12539-020-00403-6.

[9] S. Hira, A. Bai, S. Hira, An automatic approach based on CNN architecture to detect Covid-19 disease from chest X-ray images, Appl. Intell. 51 (5) (2021) 2864–2889, http://dx.doi.org/10.1007/s10489-020-02010-w.

[10] Wahab A. Musleh, A.Y. Maghari, COVID-19 detection in X-ray images using CNN Algorithm, in: 2020 International Conference on Promising Electronic Technologies (ICPET), 2020, pp. 5–9, http://dx.doi.org/10.1109/ICPET51420.2020.00010.

[11] L.T. Duong, N.H. Le, T.B. Tran, V.M. Ngo, P.T. Nguyen, Detection of tuberculosis from chest X-ray images: Boosting the performance with vision transformer and transfer learning, Expert Syst. Appl. 184 (2021) 115519, http://dx.doi.org/10.1016/j.eswa.2021.115519.

[12] V. Chouhan, et al., A novel transfer learning based approach for Pneumonia detection in chest X-ray images, Appl. Sci. 10 (2) (2020) http://dx.doi.org/10.3390/app10020559.

[13] G. Zhang, Z. Yang, L. Gong, S. Jiang, L. Wang, H. Zhang, Classification of lung nodules based on CT images using squeeze-and-excitation network and aggregated residual transformations, La Radiol. Med. 125 (4) (2020) 374–383.

[14] P. Wu, X. Sun, Z. Zhao, H. Wang, S. Pan, B. Schuller, Classification of lung nodules based on deep residual networks and migration learning, Comput. Intell. Neurosci. 2020 (2020).

[15] X. Wang, B. Cao, D. Wei, J. Liu, H. Cao, Diagnosis of thyroid nodules based on lightweight residual network, in: 2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), 2021, pp. 3875–3881.

[16] M. Mujahid, F. Rustam, R. Álvarez, J. Luis Vidal Mazón, I. de la T. Díez, I. Ashraf, Pneumonia classification from X-ray images with inception-V3 and convolutional neural network, Diagnostics 12 (5) (2022) http://dx.doi.org/10.3390/diagnostics12051280.

[17] G. Kaissis, A. Ziller, J. Passerat-Palmbach, et al., End-to-end privacy preserving deep learning on multi-institutional medical imaging | Nature Machine Intelligence. https://www.nature.com/articles/s42256-021-00337-8.

[18] R. Li, C. Xiao, Y. Huang, H. Hassan, B. Huang, Deep learning applications in computed tomography images for pulmonary nodule detection and diagnosis: A review, Diagnostics 12 (2) (2022) 298.

[19] M. Toğaçar, B. Ergen, Z. Cömert, F. Özyurt, A deep feature learning model for pneumonia detection applying a combination of mRMR feature selection and machine learning models, IRBM 41 (4) (2020) 212–222, http://dx.doi.org/10.1016/j.irbm.2019.10.006.

[20] R. Alsharif, Y. Al-Issa, A.M. Alqudah, I.A. Qasmieh, W.A. Mustafa, H. Alquran, PneumoniaNet: Automated detection and classification of pediatric pneumonia using chest X-ray images and CNN approach, Electronics 10 (23) (2021) http://dx.doi.org/10.3390/electronics10232949.

[21] Data Protection Act 2018. https://www.legislation.gov.uk/ukpga/2018/12/part/2/chapter/2/enacted.

[22] Hegedűs G. Danner, M. Jelasity, Decentralized learning works: An empirical comparison of gossip learning and federated learning, J. Parallel Distrib. Comput. 148 (2021) 109–124, http://dx.doi.org/10.1016/j.jpdc.2020.10.006.

[23] T.C. Lee, N.U. Shah, A. Haack, S.L. Baxter, Clinical implementation of predictive models embedded within electronic health record systems: A systematic review, Informatics 7 (3) (2020) http://dx.doi.org/10.3390/informatics7030025.

[24] R. Kumar, et al., An integration of blockchain and AI for secure data sharing and detection of CT images for the hospitals, Comput. Med. Imaging Graph. 87 (2021) 101812, http://dx.doi.org/10.1016/j.compmedimag.2020.101812.

[25] D. Vyas, M. Han, L. Li, S. Pouriyeh, J.S. He, Integrating blockchain technology into healthcare, in: Proceedings of the 2020 ACM Southeast Conference, New York, NY, USA, 2020, pp. 197–203, http://dx.doi.org/10.1145/3374135.3385280.

[26] R. Kumar, et al., Blockchain-federated-learning and deep learning models for COVID-19 detection using CT imaging, IEEE Sens. J. 21 (14) (2021) 16301–16314, http://dx.doi.org/10.1109/JSEN.2021.3076767.

[27] Ziller D. Usynin, R. Braren, M. Makowski, D. Rueckert, G. Kaissis, Medical imaging deep learning with differential privacy, Sci. Rep. 11 (1) (2021) http://dx.doi.org/10.1038/s41598-021-93030-0.

[28] Pfitzner N. Steckhan, B. Arnrich, Federated learning in a medical context: A systematic literature review, ACM Trans. Internet Technol. 21 (2) (2021) 50:1–50:31, http://dx.doi.org/10.1145/3412357.

[29] Feki S. Ammar, Y. Kessentini, K. Muhammad, Federated learning for COVID-19 screening from chest X-ray images, Appl. Soft Comput. 106 (2021) 107330, http://dx.doi.org/10.1016/j.asoc.2021.107330.

[30] W. Zhang, et al., Dynamic-fusion-based federated learning for COVID-19 detection, IEEE Internet Things J. 8 (21) (2021) 15884–15891, http://dx.doi.org/10.1109/JIOT.2021.3056185.

[31] Bonawitz, et al., Practical secure aggregation for privacy preserving machine learning, 2017, Cryptology ePrint Archive Available: https://eprint.iacr.org/2017/281.

[32] Zhang B. Shen, A. Barnawi, S. Xi, N. Kumar, Y. Wu, FedDPGAN: Federated differentially private generative adversarial networks framework for the detection of COVID-19 pneumonia, Inf. Syst. Front. 23 (6) (2021) 1403–1415, http://dx.doi.org/10.1007/s10796-021-10144-6.

[33] A. Chowdhury, H. Kassem, N. Padoy, R. Umeton, A. Karargyris, A review of medical federated learning: applications in oncology and cancer research, in: Proceedings of the International MICCAI Brainlesion Workshop, Virtual Event, 27, 2021, Springer, Cham, Switzerland, 2022, pp. 3–24.

[34] A. Raza, K.P. Tran, L. Koehl, S. Li, Designing ecg monitoring healthcare system with federated transfer learning and explainable AI, Knowl.-Based Syst. 236 (2022) 107763.

[35] R. Tang, J. Luo, J. Qian, J. Jin, Personalized federated learning for ECG classification based on feature alignment, Secur. Commun. Netw. 2021 (2021) 6217601.

[36] T.S. Brisimi, R. Chen, T. Mela, A. Olshevsky, I.C. Paschalidis, W. Shi, Federated learning of predictive models from federated electronic health records, Int. J. Med. Inform. 112 (2018) 59–67.

[37] A. Linardos, K. Kushibar, S. Walsh, P. Gkontra, K. Lekadir, Federated learning for multi-center imaging diagnostics: A simulation study in cardiovascular disease, Sci. Rep. 12 (2022) 3551.

[38] M. Moshawrab, M. Adda, A. Bouzouane, H. Ibrahim, A. Raad, Cardiovascular events prediction using artificial intelligence models and heart rate variability, Procedia Comput. Sci. 203 (2022) 231–238.

[39] P.V. Astillo, D.G. Duguma, H. Park, J. Kim, B. Kim, I. You, Federated intelligence of anomaly detection agent in IoTMD-enabled diabetes management control system, Future Gener. Comput. Syst. 128 (2022) 395–405.

[40] J. Lo, T.Y. Timothy, D. Ma, P. Zang, J.P. Owen, Q. Zhang, R.K. Wang, M.F. Beg, A.Y. Lee, M.V. Sarunic, et al., Federated learning for microvasculature segmentation and diabetic retinopathy classification of OCT data, Ophthalmol. Sci. 1 (2021) 100069.

[41] H. Islam, A. Mosa, A federated mining approach on predicting diabetes-related complications: demonstration using realworld clinical data, in: Proceedings of the AMIA Annual Symposium San Diego, CA, USA, 30 October–3 2021, Vol. 2021, American Medical Informatics Association, Bethesda, MA, USA, 2021, p. 556.

[42] C. Nielsen, A. Tuladhar, N.D. Forkert, Springer, in: Proceedings of the International Workshop on Ophthalmic Medical Image Analysis, Singapore, vol. 22 2022, Cham, Switzerland, 2022, pp. 183–192.

[43] H. Lee, Y.J. Chai, H. Joo, K. Lee, J.Y. Hwang, S.M. Kim, S.-M. Kim, K. Kim, I.-C. Nam, H.J. Kong, et al., Federated learning for thyroid ultrasound image analysis to protect personal information: Validation study in a real health care environment, JMIR Med. Inform. 9 (2021) e25869.

[44] P. Wang, C. Shen, H.R. Roth, D. Yang, D. Xu, M. Oda, K. Misawa, P.-T. Chen, K.-L. Liu, K. Mori, et al., Automated pancreas segmentation using multi-institutional collaborative deep learning, in: Proceedings of the Domain Adaptation and Representation Transfer, and Distributed and Collaborative Learning, MICCAI 2020, Lima, Peru, 4–8 2020, Springer, Cham, Switzerland, 2020, pp. 192–200.

[45] D. Kermany, K. Zhang, M. Goldbaum, Labeled optical coherence tomography (oct) and chest x-ray images for classification, 2018, http://dx.doi.org/10.17632/rscbjbr9sj.2.