Federated Learning-based secure Electronic Health Record sharing scheme in Medical Informatics

Mikail Mohammed Salim, and Jong Hyuk Park*

***Abstract*— Medical Cyber-Physical Systems support the mobility of electronic health records data for clinical research to accelerate new scientific discoveries. Artificial Intelligence improves medical informatics, but current centralized data training and insecure data storage management techniques expose private medical data to unauthorized foreign entities. In this paper, a Federated Learning-based Electronic Health Record sharing scheme is proposed for Medical Informatics to preserve patient data privacy. A decentralized Federated Learning-based Convolutional Neural Network model trains data locally in the hospital and stores results in a private InterPlanetary File System. A secondary global model is trained at the research center using the local models. Private IPFS secures all medical data stored locally in the hospital. The novelty of this study resides in securing valuable hospital biomedical data useful for clinical research organizations. Blockchain and smart contracts enable patients to negotiate with external entities for rewards in exchange for their data. Evaluation results demonstrate that the decentralized CNN model performs better in accuracy, sensitivity, and specificity, similar to the traditional centralized model. The performance of the Private IPFS exceeds the Blockchain-based IPFS based on file upload and download time. The scheme is suitable for promoting a secure and privacy-friendly environment for sharing data with clinical research centers for biomedical research.**

***Index Terms*—Artificial Intelligence, Bioinformatics, Data Privacy, File Systems.**

I. Introduction

Hospitals store and manage digital versions of Electronic Health Records (EHR) [1, 2] that contain the medical treatment history of all their patients. Each individual's EHR focuses on the patient's overall health, ensuring healthcare professionals are aware of any life-threatening allergies before Disease Control and Prevention has urged enterprise software development vendors that provide EHR solutions to hospitals to accommodate vaccination records when patients receive their immunization drugs [3]. prescribing any drug or surgical procedure. The Center for In Medical Cyber-Physical Systems, data collected from smart medical devices in the hospital transmit data automatically on the Cloud for storage. Artificial Intelligence (AI) based algorithms learn on local data and provide solutions for illnesses such as Valvular Heart Disease [4], Malaria [5], Parkinson's [6], Covid-19 [7, 8], and Alzheimer's Disease [9]. Collected medical data and results achieved from AI algorithms contribute to helping patients get a definitive diagnosis, treatment, and care [10].

EHR records stored in cloud servers are exposed to cyberattacks resulting in private patient data being disclosed to unauthorized entities, such as insurance organizations and fraudulent marketing establishments [11]. In 2019, there were 519 data breaches in healthcare service providers affecting 41.4 million American citizens [12]. Various AI models train their data using centralized and unsecured servers vulnerable to model input manipulation using adversarial attacks [13]. Poor model training affects hospitals that share valuable medical data required for clinical research for ailments such as Covid-19 [14]. Furthermore, patients do not maintain control over their health record access control and rely on the hospital to securely store and manage their data. Therefore, the next generation Medical Cyber-Physical Systems require a secure and privacy-preserving scheme to safeguard patient data storage and protect the AI model from manipulation.

The proposed Secure EHR Sharing scheme implements Federated Learning and functions as a decentralized learning platform. Instead of aggregating the entire data at a central location and then training the model, Federated Learning implements a shared global model further distributed to separate clients for further training. All clients train their local models and upload results to the central model for final training. In our scheme, a clinical research center functions as the site where the global model is trained and the hospital where the local is trained. Using FedAvg [15], the Federated Learning model operates with non-identical distribution data (non-IID) and requires all local clients to share their models. The data security challenge is resolved by storing all data in InterPlanetary File Systems (IPFS) in the proposed scheme. A private IPFS comprising all hospital systems ensures all patient information is available locally and not accessible by intruders. External entities such as insurance organizations and other hospitals require approval from patients using smart contracts. Patients can negotiate a better reward for them in exchange for their valuable data. Records of all authentication approvals, data access time, and data storage address are securely stored in

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (NRF-2019R1A2B5B0107041613).

Mikail Mohammed Salim, is with Seoul National University of Science and Technology, Seoul, Republic of Korea. (e-mail: mikail@seoultech.ac.kr).

Jong Hyuk Park, is with Seoul National University of Science and Technology, Seoul, Republic of Korea. (e-mail: jhpark1@seoultech.ac.kr).

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2022.3174823, IEEE Journal of Biomedical and Health Informatics

a Consortium blockchain network.

A. Related Studies

Recent studies for healthcare data security have focused on securing electronic medical records using IPFS and Blockchain technology. A ciphertext policy attribute-based encryption system in [16] protects data from unauthorized access and is stored on the IPFS. Public Blockchain records the IPFS hash and logs when data is accessed. An EHR sharing framework in [17] combined both Blockchain and IPFS for data sharing on the Mobile Cloud. Smart Contracts enable the detection of malicious access to EHRs and are accessible by patients to control user access. A public health record management method in [18] combined Blockchain and IPFS systems for a secure and traceable data storage system. A reputation-based re-encryption oracle fetches, stores, and shares patient data. A Multiparty consent management scheme in [19] relies on the guardians, healthcare regulatory agencies, and hospitals to grant user access to private patient data. IPFS system stores data, and reputation-based oracles resolve computation challenges. A medical record management scheme in [20] implements Hyperledger based Blockchain for secure data storage. Data are encrypted using the Simon technique, and upon decryption, Deep Learning based Variational AutoEncoder detects the existence of diseases. A variation-aware Federated Learning model with image privacy preservation in [21] addresses the cross-client variation problem in medical image data training.

In the above-mentioned recent research, medical health record sharing schemes secure patient private data. However, there are problems in limitations in these works, and critical research questions must be addressed for a Secure EHR Sharing scheme for medical informatics. Firstly, how to guarantee that systems part of a Public IPFS network [16-19] are online at all times. Data saved in the IPFS system is subject to being deleted if the systems where objects are stored cannot communicate. Secondly, to ensure access control of private medical records is managed by the patient alone [19]. Beyond hospitals, Health regulatory bodies should not be able to circumvent a patient's right to choose or refuse to share their data. Lastly, an AI-based model is exposed to adversarial attacks [20, 21] and requires a secure data backup for future model corrections.

An EHR sharing scheme requires design principles to satisfy the new service requirements of a secure Medical Cyber-Physical system. The proposed Secure EHR Sharing scheme takes into consideration 4 key considerations. Data Privacy, where the patient data is not exposed to unauthorized entities during model training. Data Integrity addresses storing valuable patient data securely in a decentralized storage system to avoid a single point of failure vulnerability. Access Control measures ensure each patient has complete control over sharing their private with external entities such as insurance organizations. Lastly, Data Availability is essential to recall model training data when inaccurate local model updates are sent to the clinical research centers.

B. Research Contributions

In this paper, a new Secure EHR Sharing Scheme benefits from the decentralized Federated Learning AI model, a private and public IPFS, Consortium Blockchain network, and Smart Contract technology. The main contributions of this paper are as follows:

1) A decentralized CNN-based Federated Learning model trains a global model at research centers and a local model at hospitals for Covid-19 detection.
2) Data collected from medical devices and the results of the local model are stored in Private IPFS managed on the local hospital network.
3) A cluster of trust enables the clinical research center, to directly access all local model results on the Private IPFS.
4) Consortium Blockchain network records all global model results in a Public IPFS, which a patient can use to share their medical record with third-party entities using smart contracts.
5) Evaluation of the results demonstrates similar accuracy, sensitivity, and specificity of the decentralized CNN model compared with a centralized CNN model.
6) Private IPFS outperforms Blockchain-based IPFS based on the file upload and downloads time.

The remainder of this paper is as follows. Section II presents the Federated Learning-based secure Electronic Health Record sharing scheme. Section III presents the evaluation performance of the Federated Learning model, IPFS, and the Blockchain network. Finally, Section IV concludes the paper.

II. Proposed Scheme

This paper aims to secure a patient's health records stored in Medical Informatics while maintaining user privacy and data integrity. In this section, data collected from the hospital devices are trained, and its results are stored locally. The InterPlanetary File System (IPFS) securely stores private data, and the Blockchain network stores the hash of data stored in the IPFS. Federated Learning enables training of patient data both locally and globally. Smart Contracts allow patients to control user access to their health records.

A. Secure EHR Management Scheme

Fig.1 illustrates the environment overview of the proposed Secure EHR Management Scheme. In this paper, the proposed scheme includes a medical cyber-physical system that consists of a hospital, a research center, patients present in the hospital for covid detection, and external entities such as insurance companies and foreign research labs. The hospital screens all patients for possible covid-19 infections. The Federated Learning model trains data on two ends, the hospital and the research center. A local model is trained within the hospital premises on the local tower, and results are stored in a private IPFS. A global model is managed by the research center that collects results from various hospitals to improve further and build a reliable model for the Covid-19 database. Two IPFS are present, a private and a globally distributed network. Private IPFS maintains the privacy of patient data and connects systems within the hospital premises and the research center. A cluster of trust includes the hospital medical staff and the research center. A global distributed network-based IPFS shares data with all nodes in the vast network. Smart Contracts enable third-party entities such as the research center to gain private patient data. A Consortium Blockchain-based network environment

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2022.3174823, IEEE Journal of Biomedical and Health Informatics

consists of hospitals and the research center. The hash address of the data stored in IPFS is stored securely as a transaction.

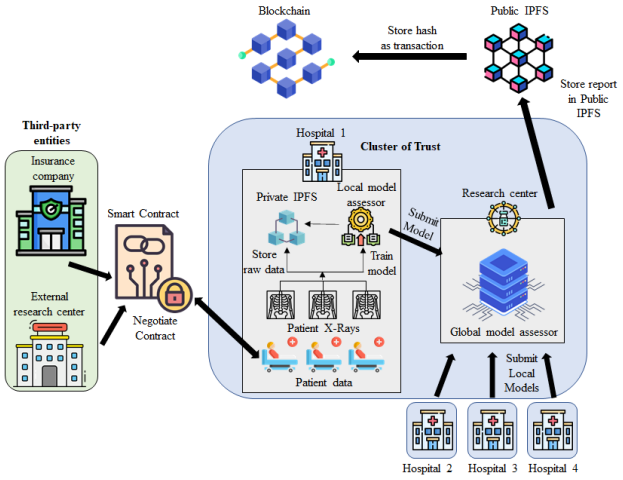

Fig. 1. Secure EHR Sharing Scheme Overview

Private IPFS include the systems locally part of the hospital systems. A private IPFS has six objectives in this scheme. Firstly, locally store all raw data collected from medical devices, ensuring no foreign entity has access to private medical data. The raw data can be further reused for other medical diagnosis systems. Secondly, store results obtained from training the local federated learning-based model to assist the local medical staff makes further informed decisions. Thirdly, ensure user access control to all elements part of the Cluster of Trust, including all relevant medical staff in the hospital. Next, prevent all foreign entities outside the trust cluster from gaining access to data to which the patient has not been given access, such as foreign research centers, other hospitals, insurance companies, and private research centers. Next, as centralized storage points are prone to the single point of failure vulnerability, a distributed file-sharing system prevents a malicious attacker from gaining access to the entire data as each data is shared in various small objects. Finally, as Public IPFS is prone to data unavailability due to public devices having the insufficient motivation to remain online, a private IPFS ensures data longevity and availability as all systems are part of the local hospital and are always available online.

Global IPFS similarly include connected peer-to-peer systems, but they are external to the hospital system. Global IPFS has the following four objectives. Firstly, all hospitals store results obtained from their respective local federated learning models. Secondly, a secure file storage system is realized as no one server stores the data, making it difficult for attackers to steal user data. Third, each file maintains a unique hash which serves as a fingerprint and ensures data integrity. Any external entity that gains access to data can verify using the hash value of information is tampered with during transmission. Finally, an external entity such as a private research center or insurance company can check for previous versions of all modified data. Once user access is secured, all previous and current versions of data are accessible to all concerned entities.

Federated Learning-based distributed learning ensures that the artificial intelligence models are not trained on centralized databases such as private Cloud networks. In this paper, Federated Learning has the following objectives. Firstly, patient data collected for identifying Covid-19 infected patients is based on a dataset using CT scans and Chest X-ray images. Secondly, a Deep Convolutional Neural Network (D-CNN) extracts features, takes input as images, and provides the output of Covid-19 infection. Next, all data are collected locally from hospital medical systems, and the model is trained locally using data centers present within the hospital. Trained models results are stored in the local IPFS to ensure user privacy agreements are not violated. Finally, the external IPFS is updated with the model's recent results based on user access permissions.

A Consortium-based Blockchain network connects the hospitals and the research center within their cluster of trust. Each hospital and the research center forms individual groups. Cyber attackers' data manipulation attempts are prevented using the block verification system where invalid blocks with incorrect hash are discarded. The Consortium-based Blockchain network has the following objectives. A Consortium based Blockchain gives access to the research center and registered hospitals the right to publish and access transactions. External entities provided user access have only read-only access, preventing them from submitting data as a transaction. The research center improves the global model and stores it in the public IPFS. A hash address serves as the fingerprint of each data stored as a transaction in the blockchain network. Data is shared across all authorized medical institutes, including the local hospital and the research center, alleviating the challenge of data unavailability. The research center accesses the hash address stored in the blockchain network and directly uploads the data from the respective local IPFS's. Immutable ledgers ensure neighboring blocks discard all data tampering attempts due to the change in the hash value.

Smart Contracts enable a patient to give consent or deny privileges to external entities to access private data. In this paper, we assume that the local medical system in the hospital is secure. Each patient can access their data freely using their mobile device and can refuse or accept smart contracts. An insurance organization may agree to subsidize a patient's hospital bill in exchange for data, or a foreign research center may promise future vaccine support in exchange for data access. In this paper, we assume that all third-party entities requesting access to user data are genuine and the patient has the means to differentiate between valid organizations and fraudulent organizations. In the proposed scheme, Smart Contracts serve four objectives. Firstly, all third-party entities are required to authenticate with patients for user access rights to private medical records. Secondly, Smart Contracts validate foreign entities with access to patient data with a version limit, i.e., the entity will either have full access to all versions of future updated EHR or only the current available EHRs. Thirdly, patients gain the right to negotiate for possible future experimental medical treatments from trusted research centers or seek financial subsidies for hospital treatments from insurance organizations. Lastly, a successful agreement to the smart contract enables a third-party entity to receive the

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2022.3174823, IEEE Journal of Biomedical and Health Informatics

blockchain transaction address pointing to the hash address of the file stored in the global IPFS.

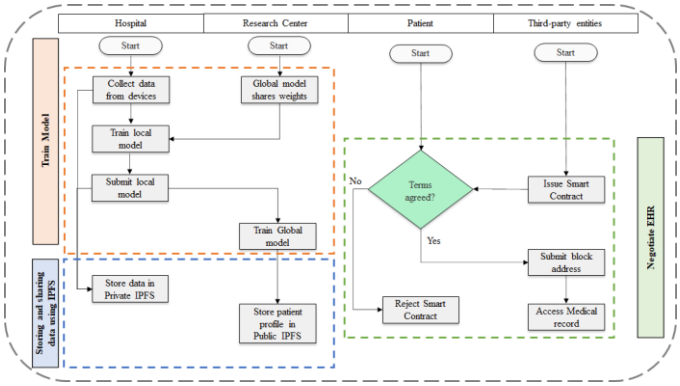

Fig. 2. Secure EHR Scheme process-flow

The Secure EHR scheme initiates, as illustrated in Fig. 2, when the patient is connected with the medical equipment that transmits data to the local private IPFS. Simultaneously, the Federated Learning model begins training the local model using X-ray and CT scans acquired from the medical devices. The scheme process flow initiates with the local model is trained from the acquired patient data. The results of the local model are stored in the private IPFS. Being within the Cluster of Trust, the research center gains immediate access to the data without sending a request message to the hospitals. The global model is trained using data collected from all the hospitals. The results are stored in a global IPFS, accessible by all medical centers being members of the consortium blockchain members.

The patient has access rights to the EHR stored in the public IPFS. Using Smart Contracts, the patient issues access rights to third-party entities and transfers the transaction address. Revocation or expiry of the Smart Contract invalidates the authentication of the entity accessing the data. Hash addresses stored in blocks consisting of future updates of the EHR file versions stored in blocks are not shared with other entities with expired Smart Contracts.

The Secure EHR scheme implements a four-phase approach, 1) Train Model, 2) Storing and share data using IPFS, 3) Store IPFS records, and 4) Negotiate EHR.

B. Train Model

In this paper, Federated Learning implements the D-CNN algorithm for decentralized learning. X-ray images are classified based on the detection of active Covid-19 using the FedAvg algorithm. The training of the model is based on several rounds (n) where the model begins with weights (wt^0) and the number of hospitals (hsp) and each hsp has several images (img_{hsp}) of the patient. The global model shares the weights (wt^{n-1}), where each hospital $(hsp_1, hsp_2, hsp_3, \ldots\ldots hsp_n]$ has access to $[wt_1, wt_2, wt_3, \ldots\ldots wt_n]$. Weights are shared randomly with a subset of hospitals (h_{st}). As shown in Algorithm 1, each hospital $hsp \in h_{st}$ trains the local data based on the X-ray images. Training is done in batches (bt) using the mini-batch Stochastic Gradient Descent (SGD). The number of epochs to run SGD on data points is termed as Ep.

Algorithm 1: Local model training

Input: wt_n, training data, n
Procedure: local model update ($\mathrm{wt_n}$)
  $\mathrm{wt_1} \leftarrow \mathrm{wt_n}$
  distribute data into bt
  for each ep run Stochastic Gradient Descent
    for each bt_n in bt do
      perform gradient update
      update local model
    end for
  end for
  return $\mathrm{wt_1}$
end procedure

Upon completion of the training of the model, each h_{st} uploads its local model to the global model/research center, $wt^t_{hsp}, hsp \in h_{st}$. As shown in Algorithm 2, the research center averages the local models received from the h_{st} and updates the global model parameters.

Algorithm 2: Global model training

Input: Fed_n, l_{fun} (Federated round, Loss function)
Procedure: Aggregate $\mathrm{wt_n}$ ($\mathrm{wt_1}, \mathrm{wt_2}, \mathrm{wt_3} \ldots\ldots \mathrm{wt_n}$)
  Begin global model
  for each $\mathrm{Fed_n}$ fed=1, 2, 3……..$\mathrm{fed_n}$ do
    select random weights of hsp $\mathrm{wt_n}$
      for each $hsp \in h_{st}$
        update global model
      end for
      FedAvg (Aggregate hsp updates)
  end for
  Store in IPFS
end procedure

The above single round of D-CNN is part of several rounds designed for federated learning. The research center repeatedly sends the weights $\mathrm{w^{n-1}}$ and as each new hospital becomes part of the secure EHR scheme, the number of hsp_{n+1} increases.

The CNN model uses the two architectures, ResNet50 and VGG16. The scheme includes a prior-trained model. The classification head consists of dropout for the two CNN architectures with 256 and 64 units connected layers. The corresponding final layer consists of 2 units with softmax activation optimized using the category cross-entropy function. For the representation of the local hospital side model, the scheme shows one hospital training the local model in this paper. The remaining hospitals share the same training in D-CNN-based architecture. The global model managed by the research center initiates Federated Learning by distributing the model to the hospital for local training. As the hospital submits its trained model, the result is used to improve the global model.

C. Storing and sharing data using IPFS

The design of a Private and Public IPFS is similar except that in Private IPFS, only systems present on-site in the hospital are allowed to connect. Doctors, nurses, and administrators

have direct access to EHR, being a member of the cluster of trust. Public IPFS are open access to all systems available in the public sphere. The objective of the IPFS is to be a safe and privacy-friendly storage system for sensitive EHRs. Each EHR is stored in the format of a PDF as an object. Each object is 256 kb in storage space and therefore is lightweight. The EHR is shared across several systems, ensuring that no single system is vulnerable to a lack of storage space. Furthermore, each file is represented by an SHA-256-bit hash address that serves as a pointer to assemble the file and check the integrity of the aggregated EHR. Vast amounts of files are possible to store on various systems local to a single hospital. There are three types of communication with IPFS in this scheme,

1) Machine to IPFS communication: Machines part of the local hospital system transmit raw data and store in the local IPFS network for future analysis.
2) Federated Learning local model to IPFS communication: Completing local data training requires storing the results in the private IPFS for local access in the hospital network.
3) IPFS to Research center: The Research center has direct access to all learned data from the hospital to improve the global model. The research center is not required to request data from hospitals but instead access the premade available data in the private IPFS of all hospitals.
4) Smart Contract to Blockchain communication: The Public IPFS requires aggregating data for valid users authorized by the smart contract. A valid smart contract enables searching for the hash address stored in the blockchain network. A certified member requests for the data and the public IPFS retrieves the information.

A key benefit of keeping two separate IPFS, private and public, prevents a third-party entity from accessing the raw data used to train the local and global model. An authorized entity can access only the medical report file information as the research center and hospital allow.

D. Store IPFS Records

In this scheme, the Blockchain network is not used to store data as Medical records multiply and introduce scalability problems in the network. Therefore, all data are stored in the private and public IPFS. The Consortium Blockchain network is responsible for maintaining a record of the location address where each EHR is recorded. The medical research center stores data in the Public IPFS. Each stored file has a hash address which is stored in the Blockchain network as a transaction. Stored transaction IDs are maintained by the research center, the local hospital, and the concerned patient. All nodes are verified by their neighboring nodes, and an intrusion attempt to manipulate data results in the block being rejected. Therefore, ensuring data integrity. The immutable ledger maintains a transaction record of each access authorization and invalidation using smart contracts. The description includes the list of all entities granted privileges to access the record by the patient. The information stored consists of the time constraints attached to the smart contract-based agreement between the third-party entity and the patient.

E. Negotiate EHR

The proposed Secure EHR scheme prevents any third-party entity from accessing private patient records. However, in the patient's interests and based on their personal decision, Smart Contracts enable sharing of health records with an external entity. This paper assumes that the patient has a mobile device to receive and accept or refuse a smart contract offer. The process initiates when an external entity such as a foreign medical research center or an insurance company requests the user to share data based on the provided agreement. The patient has the means to accept, reject, or negotiate the terms of agreement for monetary or medical benefit. Accepting the contract requires the patient to share the blockchain transaction address for the external entity to access the EHR. Furthermore, the period of access and user access restriction on future versions of the file can be stated in the smart contract. The blockchain network receives the smart contract and records the details of the agreement in a block. The authorized user gives the hash address, and the file is downloaded using the Public IPFS's GET command.

III. Analysis

We analyze the proposed Secure EHR scheme using Ubuntu 18.04 and i7 processor with 32 Gb RAM. The evaluation is based on the performance of the Federated Learning model, the IPFS file upload and download speed, and the blockchain transaction upload and transaction download speed. The Federated Learning model uses the CNN training model using RESNET50 and VGG16 architectures. Dataset for training contains Chest X-Ray images of 76 patients and is equally distributed with 32 healthy and 32 Covid-19 infected patients [22, 23]. Simulations are performed using $hsp = 4$ hospitals. The IPFS networks are installed using the go-ipfs version 0.10.0 [24] and the Blockchain network is designed using Hyperledger Fabric Blockchain.

A. D-CNN based Federated Learning

Evaluation of the model is based on the accuracy for identifying Covid-19 patients and further preventing the spread of the disease. The sensitivity metric reflects the true positivity rate in accurately detecting infected patients, and the specificity metric outlines the true negatives, i.e., correctly identifying healthy patients. Table 1 lists the dataset specifications.

TABLE I
DATASET SPECIFICATIONS

Description	**Parameters**	
Datasets	Covid X-ray images [22]	Normal X-ray images [23]
Hospitals	4	
X-Ray images	108	
Patients	76	
Dataset split	80% (Training set) - 20% (Test set)	
Training set	55 patients	
	76 Covid images	76 healthy images
Test set	21 patients	
	32 Covid images	32 healthy images

As illustrated in Fig 3, the decentralized Federated Learning model (DL) is compared with Centralized Learning (CL) trained using the traditional machine learning method. The objective of the comparison is to show the increase in the accuracy of the model using decentralized training without

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2022.3174823, IEEE Journal of Biomedical and Health Informatics

sharing the data on a centralized system. Each round out of 4 is trained using the VGG16 and the RESNET50 based CNN architecture. The Federated-based decentralized model is presented based on an individual round that represents a local model shared by a hospital. Instead, the centralized model is based on the number of data-sharing epochs where each epoch indicates a training pass over the entire data. The accuracy of the DL improves as the number of rounds increases at the cost of training time.

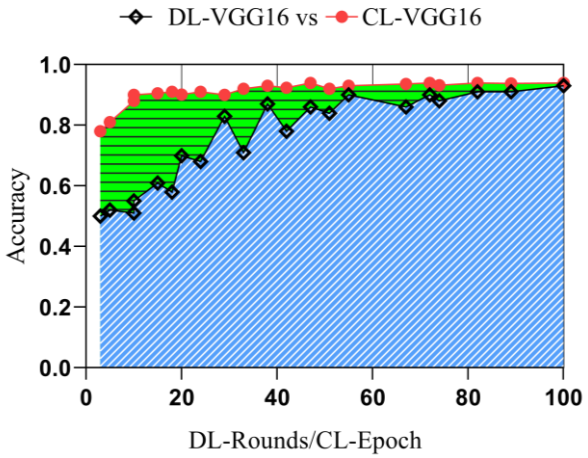

Fig. 3. Accuracy of DL-VGG16 and CL-VGG16 models

In Fig 4, we observe that as the training reaches around 38, it matches the accuracy produced by the centralized model using the same VGG16. In Fig 3(b), the Resnet50 has similar results where the model reaches the accuracy from the centralized training model but took 74 rounds. Therefore, from the results, with increased rounds, the DL model is as accurate as the centralized model and without compromising the privacy of EHR. A noticeable delay in DL model training is observed compared with the CL-CNN model.

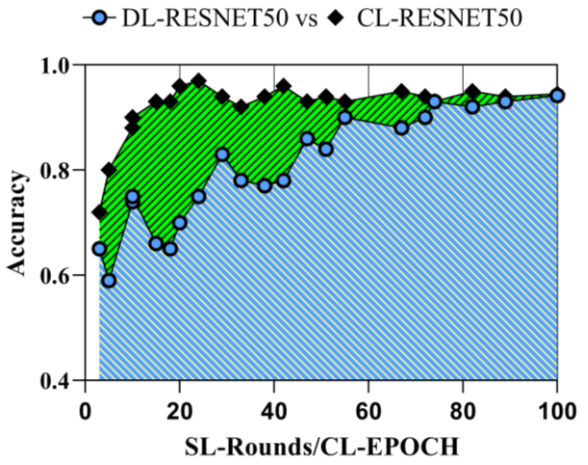

Fig. 4. Accuracy comparison between DL-RESNET50 and CL-RESNET50 models

The final measurement of accuracy, sensitivity, and specificity is measured by observing the last round of the decentralized Federated Learning. The centralized models are measured by observing the results obtained from the final epoch. Table 2 outlines the results obtained for both decentralized and centralized models.

TABLE II
ACCURACY, SENSITIVITY, AND SPECIFICITY

Process	Results		
	Accuracy	Sensitivity	Specificity
DL-VGG16	93.18	94.86	92.89
CL-VGG16	92.34	95.44	92.62
DL-RESNET50	94.28	96.37	93.74
CL-RESNET50	94.48	95.1	93.93

The final round and the final epoch results indicate the decentralized VGG16 and RESNET50 model performances. The accuracy, sensitivity, and specificity of Federated Learning (DL) are comparable with the centralized models.

B. IPFS and Blockchain-based IPFS

Evaluation of the performance of the IPFS system is based on the file upload and download performance compared with a Blockchain-based IPFS. The Federated Learning model requires quick access to all data stored in the IPFS for the global model training. To ensure no delay in preparing the final model, local hospitals must quickly upload results obtained from the local model. The private IPFS network is designed using a system running Ubuntu 18.04 and installed the go-ipfs version 0.10.0. The local node and one other system are included as part of an IPFS swarm. The Blockchain network is deployed using the Ethereum Blockchain. To simulate the performance of the file upload and download, we measured the performance between one client node (hospital) and one bootstrap node.

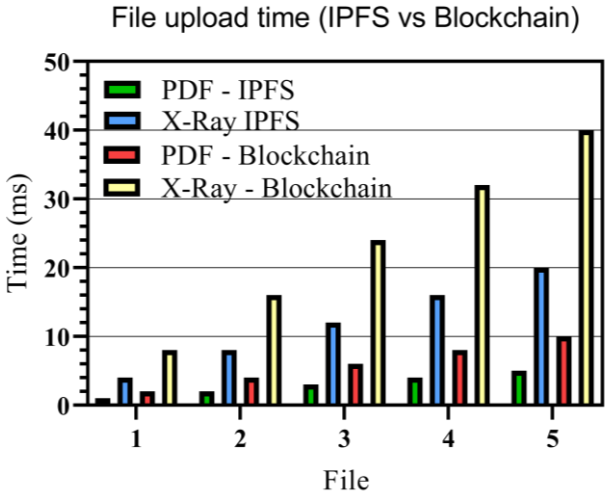

Fig. 5. X-Ray and PDF file upload time comparison

Fig. 5 illustrates the performance impact when uploading files to the Private IPFS network. 10 Files are uploaded, each of varying size in Megabytes (MB), and the time is measured in milliseconds (ms). Five small-size files ranging between 5 – 25 MB indicate PDFs uploaded containing primarily text-based records. In contrast, the larger files, 20 MB – 60 MB, represent images belonging to patients to mimic X-rays and uploaded on the IPFS network. The performance is compared with a blockchain-based IPFS and performs upload and download operations.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2022.3174823, IEEE Journal of Biomedical and Health Informatics

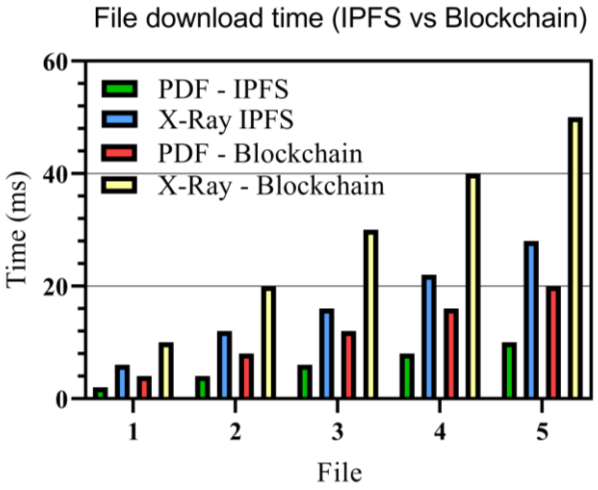

Fig. 6. X-Ray and PDF file download time comparison

In Fig. 6, uploading five small PDF files, and five large image files, the time consumed by the Private IPFS is less than the Blockchain-based IPFS. The file is uploaded by the local node (hospital), and the other node (research center) accesses the file. The Blockchain-based IPFS takes more time to upload and download the files as it depends on completing the IPFS file processes. The results indicate that the private IPFS system is ideal for sharing EHR between the local hospital and the research center.

C. Storing IPFS hash on Blockchain

We evaluate the time taken to upload a complete EHR record on the Blockchain network by the research center and compare it with storing a hash address pointing to the file stored in the public IPFS network. For evaluation, we store five SHA-256 hash addresses pointing to five different medical records. Furthermore, we consider a single EHR record of 64 bytes and upload five pdf files to store on the blockchain network. Each file is stored in a block and downloaded to measure the performance impact.

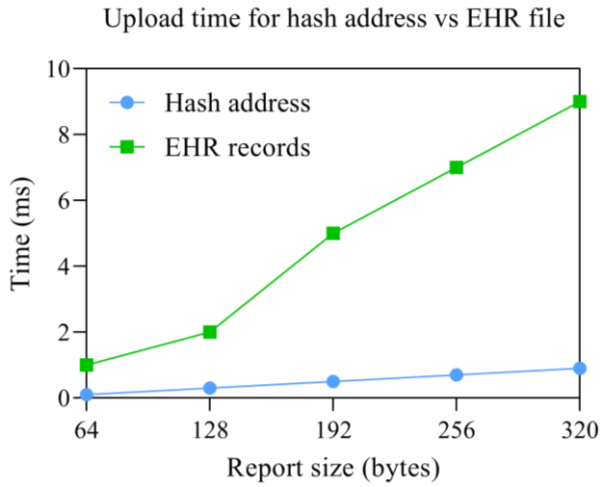

Fig. 7. Upload time for uploading IPFS hash address and EHR record in Blockchain

As shown in Fig. 7, the time required to upload the hash address is 0.9 ms. for five 320-byte size files, or 90% less compared with a full EHR upload that required 9 ms. on the Blockchain and increases proportionately as the count of files increases. We observe that as file byte size increases, uploading a file on the Blockchain introduces increased latency presenting network congestion challenges. We determine that even if the file size is reduced, the reduced size of a hash address outperforms a full EHR storage. As EHR records grow, IPFS is ideal for data storage compared to Blockchain for Healthcare based systems. Downloading a hash address requires less time and is thus suitable for third-party entities. However, downloading large block files takes more time, so a private IPFS is ideal for global model training.

D. Discussion

The field of Bioinformatics benefits from obtaining accurate and secure data from local hospitals to track the spread of contagious diseases among the population. Table 3 presents the comparison between related studies and the proposed scheme based on fulfilling the key design principles for a secure EHR sharing scheme. The proposed scheme enables clinical research centers to collect data from hospitals using Federated Learning. Data Privacy concerns are maintained by training data locally within hospitals, and only sharing gradients with the clinical research center.

TABLE III
COMPARISON OF THE PROPOSED SCHEME WITH EXISTING STUDIES

Existing Research	Key areas of Consideration			
	Privacy	Integrity	Access Control	Availability
Sun et al. [14]	×	✓	×	✓
Nguyen et al. [15]	×	✓	×	✓
Madine et al. [16]	×	✓	×	✓
Madine et al. [17]	×	✓	✓	✓
Sammeta et al. [18]	×	✓	✓	✓
Yan et al. [19]	✓	×	✓	×
Proposed Scheme	✓	✓	✓	✓

Data collected from medical devices are stored in a private and decentralized IPFS that ensures data integrity. Attackers are required to infect multiple systems to manipulate stored data and affect the local model training. Patients maintain access control using Smart Contracts and authorize entities that benefit them, such as financial incentives from insurance organizations, and experimental medical treatments from external, research centers and hospitals. Availability of data enables a hospital to retrain local models during an event an uploaded model is intercepted by an attacker. Storing raw data collected from medical devices enables a hospital to retrain data. Furthermore, local backups stored in Private IPFS prevent data theft.

IV. CONCLUSION

This paper presented a privacy-friendly and secure EHR scheme for Medical Cyber-Physical Systems. A Federated Learning-based decentralized artificial intelligence model trains data locally in the hospitals and globally at the research center. Each local hospital improves the local model and stores results on the private IPFS, ensuring that sensitive medical

records are not lost. A secure Consortium Blockchain-based network securely records all EHR hash addresses stored in the IPFS. Smart Contracts enable each patient to refuse or allow third-party entity access to private EHR. The evaluation of the scheme demonstrates that the Federated Learning model performs well in accuracy, sensitivity, and specificity compared to the traditional centralized model. The Private IPFS outperforms Blockchain-based IPFS based on the file upload and downloads time. Finally, we evaluated the time required to store and download files on the Blockchain network and observed that storing the IPFS hash address is more suitable than keeping a complete EHR on the blockchain network. The limitations of the proposed scheme are twofold, the vulnerabilities of hospital systems exposed to cyberattacks, and the scalability of Blockchain to store and manage medical big data. The future scope of the research includes addressing application layer securities and using sidechains for scalability concerns.

ACKNOWLEDGMENT

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (NRF-2019R1A2B5B0107041613).

REFERENCES

[1] J. W. Li, Y. C. Chang, M. X. Xu, and D. Y. Huang, "A Health Management Service with Beacon-Based Identification for Preventive Elderly Care," *Journal of Information Processing Systems*, vol. 16, pp. 648-662, Jun, 2020

[2] A. R. Javed, M. U. Datwar, M. O. Beg, M. Asim, T. Baker, and H. Tawfik, "A collaborative healthcare framework for shared healthcare plan with ambient intelligence," *Human-centric Computing and Information Sciences*, vol. 10, pp. 1-21, Sept 2020

[3] "How top EHR vendors are prepping their systems for COVID-19 vaccines", 2020, [online] Available: https://www.healthcareitnews.com/news/how-top-ehr-vendors-are-prepping-their-systems-covid-19-vaccines

[4] Y. S. Su, T. J. Ding, and M. Y. Chen, "Deep Learning Methods in Internet of Medical Things for Valvular Heart Disease Screening System," *IEEE Internet of Things Journal*, Jan 2021

[5] P. A. Pattanaik, M. Mittal, and M.Z. Khan, "Unsupervised Deep Learning CAD Scheme for the Detection of Malaria in Blood Smear Microscopic Images," *IEEE Access*, vol. 8, pp. 94936 – 94946, May 2020

[6] G. Nagasubramanian, and M. Sankayya, "Multi-Variate vocal data analysis for Detection of Parkinson disease using Deep Learning," *Neural Computing and Applications,* vol. 33, pp. 4849 4864, Aug 2020

[7] C. Zhou, J. Song, S. Zhou, Z. Zhang, and J. Xing, "COVID-19 Detection Based on Image Regrouping and Resnet-SVM Using Chest X-Ray Images," *IEEE Access*, vol. 9, pp. 81902 – 81912, June 2021

[8] F. Hao, and D.S. Park, "CoNavigator: A Framework of FCA-Based Novel Coronavirus COVID-19 Domain Knowledge Navigation," *Human-centric Computing and Information Sciences,* vol. 11, 2021

[9] H. Nguyen, and N. N. Chu, "An Introduction to Deep Learning Research for Alzheimer's Disease," *IEEE Consumer Electronics Magazine*, vol. 10, pp. 72 – 75, May 2021

[10] A.K. Sangaiah, D.V. Medhane, T. Han, M.S. Hossain, and G. Muhammad," Enforcing Position-Based Confidentiality With Machine Learning Paradigm Through Mobile Edge Computing in Real-Time Industrial Informatics," *IEEE Transactions on Industrial Informatics*, vol. 15, pp. 4189-4196, Feb 2019

[11] J. P. Gutierrez, and K. Lee, "High Rate Denial-of-Service Attack Detection System for Cloud Environment Using Flume and Spark," *Journal of Information Processing Systems*, vol. 17, pp. 675 – 689, Aug 2021

[12] F. Luh, and Y. Yen, "Cybersecurity in Science and Medicine: Threats and Challenges," *Trends in Biotechnology, vol. 38, pp. 825 – 828, Aug 2020*

[13] A. Rehman , M. S. Hossain, N. A. Alrajeh, and F. Alsolami, "Adversarial Examples—Security Threats to COVID-19 Deep Learning Systems in Medical IoT Devices," *IEEE Internet of Things Journal*, vol. 8, pp. 9603 – 9610, Aug 2020

[14] A.K. Sangaiah, M. Arumugam, and G.B. Bian, "An intelligent learning approach for improving ECG signal classification and arrhythmia analysis," *Artificial Intelligence in Medicine*, vol. 103, pp. 101788, Mar 2020

[15] J. Xu, B. S. Glicksberg, C. Su, P. Walker, J. Bian, and F. Wang, "Federated Learning for Healthcare Informatics," *Journal of Healthcare Informatics Research*, vol. 5, pp. 1-19, Nov 2020

[16] J. Sun, X. Yao, S. Wang, and Y. Wu, "Blockchain-Based Secure Storage and Access Scheme For Electronic Medical Records in IPFS," *IEEE Access*, vol. 8, pp. 59389 – 59401, March 2020

[17] D. C. Nguyen, P. N. Pathirana, M. Ding, A. Seneviratne, "Blockchain for Secure EHRs Sharing of Mobile Cloud Based E-Health Systems;" *IEEE Access*, vol. 7, pp. 66792 – 66806, May 2019

[18] M. M. Madine, A. A. Battah, I. Yaqoob, K. Salah, R. Jayaraman, Y. A. Hammadi., S. Pesic, and S. Ellahham, "Blockchain for Giving Patients Control Over Their Medical Records," *IEEE Access*, vol 8., pp. 193102 – 193115, Oct 2020

[19] M. M. Madine, K. Salah, R. Jayaraman, I. Yaqoob, Y. A. Hammadi, S. Ellahham, P. Calyam, "Fully Decentralized Multi-Party Consent Management for Secure Sharing of Patient Health Records," *IEEE Access*, vol. 8, pp. 225777 – 225791, Dec 2020

[20] N. Sammeta, and L. Parthiban, "Hyperledger blockchain enabled secure medical record management with deep learning-based diagnosis model," *Complex & Intelligent Systems*, pp. 1-6, Oct 2021

[21] Z. Yan, J. Wicaksana, Z. Wang, X. Yang, and K.T. Cheng, "Variation-Aware Federated Learning With Multi-Source Decentralized Medical Image Data," *IEEE Journal of Biomedical and Health Informatics*, vol. 25, pp. 2615-2628, Nov 2020

[22] 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

[23] S. Jaeger, S. Candemir, S. Antani, Y.X.J. Wang, P. X. Lu, and G. Thoma, "Two public chest X-ray datasets for computer-aided screening of pulmonary diseases," *Quantitative imaging in medicine and surgery*, vol. 4, pp. 475–477, Dec 2014

[24] "Welcome to IPFS Distributions", 2021, [Online] Available: https://dist.ipfs.io/