Wang *et al. Rare Dis Orphan Drugs J* 2023;2:22
DOI: 10.20517/rdodj.2023.16

Rare Disease and
Orphan Drugs Journal

Review

Open Access

Federated learning for rare disease detection: a survey

Jiaqi Wang, Fenglong Ma

College of Information Sciences and Technology, Pennsylvania State University, Westgate Building, University Park, PA 16802, USA.

Correspondence to: Prof. Fenglong Ma, College of Information Sciences and Technology, Pennsylvania State University, Westgate Building, University Park, PA 16802, USA. E-mail: fenglong@psu.edu

How to cite this article: Wang J, Ma F. Federated learning for rare disease detection: a survey. *Rare Dis Orphan Drugs J* 2023;2:22. https://dx.doi.org/10.20517/rdodj.2023.16

Received: 19 Jun 2023 **First Decision:** 29 Aug 2023 **Revised:** 15 Sep 2023 **Accepted:** 27 Sep 2023 **Published:** 10 Oct 2023

Academic Editor: Daniel Scherman **Copy Editor:** Dan Zhang **Production Editor:** Dan Zhang

Abstract

The detection of rare diseases utilizing advanced artificial intelligence (AI) techniques has garnered considerable attention in recent years. Numerous approaches have been proposed to detect diverse rare diseases by leveraging a range of medical data, including medical images, electronic health records, and sensory data. In order to safeguard the privacy of health data, considerable investigation has been undertaken on a novel learning paradigm known as federated learning, which has been applied to the domain of rare disease detection. Nonetheless, this nascent research direction remains in its infancy, necessitating greater scrutiny and attention. Within this survey, our primary focus lies in providing fresh perspectives, deliberating the challenges, and enumerating potential research directions concerning the application of federated learning techniques in rare disease detection. Furthermore, we provide a succinct summary of existing advancements using AI techniques for rare disease detection, as well as the utilization of federated learning within healthcare informatics. Moreover, we furnish a compilation of publicly available datasets that can be employed to validate novel federated learning algorithms for the purpose of detecting rare diseases.

Keywords: Rare disease detection, federated learning, AI for healthcare

INTRODUCTION

A rare disease is defined as a condition that affects a small number of people compared to the general population. According to the U.S. Food and Drug Administration (FDA), a disease is considered rare if it affects fewer than 200,000 individuals in the country[1]. However, different countries may have their own official definitions of a rare disease. For instance, the European Union defines a disease as rare when it affects fewer than 1 in 2,000 people[2].

Although the number of patients with rare diseases is small, the range of rare diseases is extensive, resulting in a significant overall number. There are several important facts to consider: to date, over 6,000 different rare diseases have been identified worldwide, currently affecting approximately 3.5% to 5.9% of the global population[3]. In the United States alone, there are more than 10,000 known rare diseases that affect about 1 in 10 people, totaling approximately 30 million individuals[4]. Due to the low prevalence of each disease, expertise in their diagnosis and treatment is limited, knowledge about them is scarce, care offerings are inadequate, and research is often restricted.

However, the detection of rare diseases presents considerable challenges, particularly during the early stages. Let us take pancreatic cancer as an exemplar, which stands as an exceedingly lethal form of cancer with a mere 11% overall five-year relative survival rate in the United States, the lowest among all cancer types[5]. Patients typically exhibit nonspecific symptoms, such as jaundice, fatigue, alterations in bowel habits, and indigestion, thereby complicating the differentiation from non-malignant diseases[6]. Furthermore, the identification of early-stage pancreatic disease is hampered by the absence of reliable biomarkers. While carbohydrate antigen 19-9 represents the most extensively validated biomarker for pancreatic cancer, it falls short in terms of screening accuracy and specificity[7,8]. Thus, the imperative development of innovative techniques for the detection and comprehension of rare diseases assumes profound and practical significance.

Recently, the medical field has widely adopted AI techniques to aid in the detection of rare diseases using diverse data sources, such as medical images[9-11] and electronic health records (EHR)[12-14]. However, the training of AI models, especially deep learning-based ones, typically requires a large amount of data. The challenge lies in accessing rare disease data from different institutions due to concerns over data privacy. This obstacle hinders the development of new AI techniques. Fortunately, the emergence of federated learning techniques offers a solution to address the data privacy issue[15,16]. Federated learning is a novel learning paradigm that enables collaborative training of machine learning models without sharing data with others. Although several approaches have been proposed for utilizing federated learning in solving medical problems, limited work has been dedicated to rare disease detection. However, to our best knowledge, there are no existing surveys focusing on the specific domain of rare disease detection with federated learning. Therefore, research questions arise: (1) what are the state-of-the-art research works about rare disease detection with federated learning? (2) What data and techniques are utilized in the related research works? (3) Are there any promising future research directions in this domain?

This survey primarily focuses on discussing the fundamental challenges of applying federated learning to detect rare diseases and explores potential research directions. Prior to that, we provide a brief overview of existing work in AI for rare diseases and the application of federated learning in healthcare informatics. Additionally, we present a list of commonly used datasets for rare disease detection, which can be employed for simulating experiments involving federated learning models.

METHODS

Search strategy

We identified the related research works by conducting a comprehensive search in the search engines and databases, including but not limited to Google Scholar, IEEE Xplore, ACM Digital Library, and arXiv. We used the following key words to obtain our inclusive searching results: “machine learning in rare disease detection”, “AI in rare disease detection”, “federated learning in healthcare informatics”, and “rare disease detection in federated learning”. This search was conducted up to 14 May, 2023 and all the research works which meet the inclusion criteria were under consideration.

Selection criteria

Inclusion requirements are (1) original research articles; (2) published in English; (3) full-text access; (4) significant contributions; (5) on the topics of machine learning in rare disease detection, healthcare informatics in federated learning, and rare disease detection in federated learning. Exclusions for this survey are: (1) articles not published in English; (2) full-text access not available; (3) without clear evaluation, constructive discussion, or convincing statement.

Screening process

Based on the search strategy using the databases and key words, we found 1,040,300 records. After removing the duplicated search results, considering the relevance, and filtering with the selection criteria, we have included 91 research works, datasets, web resources, and related tools for this survey. The screening process is shown in Figure 1.

MACHINE LEARNING FOR RARE DISEASE DETECTION

Rare disease

The definition of rare diseases varies across different countries. In the United States of America, a disease is considered rare if the number of affected individuals is less than 200,000[17]. Conversely, the European Union defines rare diseases as those that either affect 0.5% of the population or are life-threatening, debilitating, or chronic[18]. Rare diseases are often classified into several categories, including metabolic disorders, neuromuscular disorders, blood disorders, cardiovascular and respiratory disorders, autoimmune diseases, skin diseases, and rare neoplasms[19].

Patients with rare diseases face unique challenges compared to those with more common conditions. These challenges arise due to limited research studies, small patient communities, and a lack of attention and funding for treatment development due to economic constraints. For more detailed information on rare diseases, existing rare disease surveys can be consulted[20].

AI for rare disease detection

AI has the capability to extract hidden patterns and analyze complex relationships in data using machine learning (ML) or deep learning (DL) methods. With the increasing availability of EHR, it has become easier to access, process, and analyze patient data. However, EHR data contains diverse types of information, such as age, gender, hospital visits, diagnosis records, and lab test results. Consequently, it becomes challenging to comprehensively analyze and identify the relationships between rare diseases and these various features.

Numerous research works have focused on rare disease detection using AI. In[21], a complementary pattern augmentation framework is proposed, combining ideas from adversarial training and max-margin classification. As for adversarial training, there are several research works[22-24] using generative adversarial networks to do rare disease detection. From the *methodology perspective*, multiple deep learning techniques are also applied in rare disease detection. For instance, representation learning is utilized to detect rare

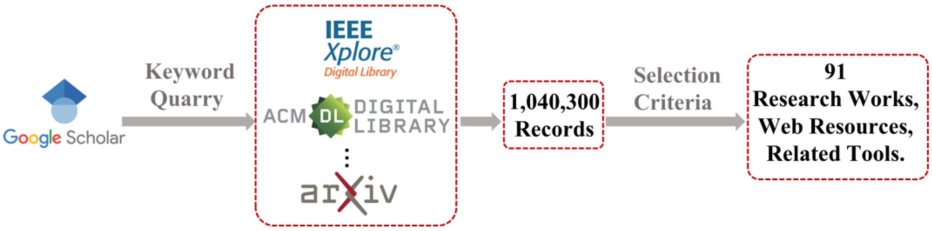

Figure 1. Overview of the screening process.

disease-associated cell subsets[25]. Few-shot learning is applied to help generate radiology reports from radiology data to support the rare disease diagnosis[26]. In[27], reinforcement learning is used to design a tool to conduct a rare disease diagnostic task using expert knowledge and clinical data with the minimum number of medical tests. In[28], natural language processing (NLP) techniques are applied to extract the relevant information of rare diseases and identify the clinical manifestations. RareBERT[29] is another research work that applies transformer-based techniques to identify rare diseases using administrative claims. In[30], a difficulty-aware meta-learning approach is used to conduct the rare disease classification with the consideration of learning task importance. In[31], a transfer learning approach is applied to obtain a machine-learning model from a sex-identification model to detect the Phospholamban mutation. Another research work[32] also employs transfer learning by training a model on a large public medical dataset and transferring it to rare disease datasets, demonstrating the algorithm's effectiveness compared to training solely on a given dataset. From the *data type perspective*, there are existing research works using EHR data[12-14], image data[9-11], and other formats of medical-related data (online searching data[33] or replies to questionnaires[34]). Finally, there are several surveys[35-46] of AI for rare diseases from different perspectives.

FEDERATED LEARNING FOR HEALTHCARE

Federated learning

Federated learning (FL) enables multiple parties to cooperate to train machine learning models without sharing data. FL is initially proposed in[47], where authors introduce a classic algorithm named FedAvg. There are several steps in FedAvg: (1) Step 1: each party trains its model using local data; (2) Step 2: active clients upload model parameters back to the server; (3) Step 3: the server aggregates the model parameters to obtain one global mode; (4) Step 4: the server distributes the global model parameters back to the active client at the next communication round as their initialization to train their local models. This process will continue until the system reaches convergence or the required communication rounds. The demonstration can be found in Figure 2.

FL has gained significant attention from researchers in academia and industry due to its advantages in distributed data utilization and data privacy preservation. Till this research, FL has been explored in multiple directions including but not limited to heterogeneity[48,49], security[50] and privacy[51], robustness[52] and fairness[53], and application domains including but not limited to healthcare[54], computer vision[55], urban computing[56], and finance[57]. Additionally, there are companies and communities providing packages, platforms, and open-source support for conducting FL industry products and research works, e.g., IBM[58], Google Cloud[59], and TensorFlow Federated[60].

Healthcare informatics in federated learning

With the development of technology, more and more healthcare-related information can be acquired, saved, and analyzed, including but not limited to electronic health records, medical images, and claim data. However, most of the patients' data are sensitive and subject to strict legal and policy regulations. FL offers a

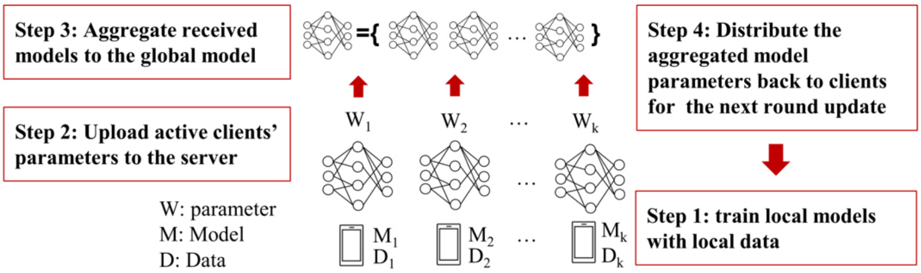
Step 3: Aggregate received models to the global model
Step 2: Upload active clients' parameters to the server
W: parameter
M: Model
D: Data
W_1
W_2
W_k
M_1 D_1
M_2 D_2
M_k D_k
Step 4: Distribute the aggregated model parameters back to clients for the next round update
Step 1: train local models with local data

Figure 2. Federated learning paradigm.

promising approach for healthcare data holders, such as hospitals or medical research institutions, to collaborate in training machine learning models without the need to share the data itself. This enables distributed data analysis while preserving data privacy.

Several surveys[61-66] have been conducted on healthcare informatics in the context of federated learning. In the survey[61], the authors discuss the existing algorithms and application of federated learning on electronic health records. Other surveys[62,63,65] provide general systematic reviews of federated learning for healthcare informatics. There is also another survey[64] that talks about the federated learning research works in clinical studies with structured medical data. Furthermore, the future directions of digital health with federated learning have been discussed as well in a recent work[66]. In the following discussion, we will explore existing applications of FL in healthcare from the perspective of data types.

Medical image

Medical imaging plays a crucial role in capturing patients' information, enabling doctors, researchers, and scientists to perform diagnoses, treatments, and research. However, different institutes may possess varying types or quantities of medical images, making collaboration essential. FL facilitates cooperation among these institutions, allowing them to work together on tasks such as COVID-19 diagnosis, cancer detection, heart disease detection, and thyroid diagnosis[67].

In[68], the cooperation of brain tumor segmentation across multiple institutes is enabled via FL. The authors propose a FL framework in[69] for detecting COVID-19 infections using Chest X-ray images from different data holders. In[70], the authors utilize FL frameworks to conduct a diagnosis of hypertrophic cardiomyopathy with magnetic resonance imaging (MRI) data.

Electronic health records

EHR contains patients' information including but not limited to hospital visits, diagnosis records, medication records, laboratory results, allergies, immunization status, and basic personal information. EHRs describe patient information in a textual format, capturing their medical history in a time-series manner, which can reveal changes in their health status over time. Deep learning techniques can be leveraged to uncover hidden relationships among the various types of data in EHRs and provide related analyses.

However, conducting FL on distributed EHR data poses significant challenges in the real world, particularly due to issues such as data imbalance and missing data. In[54], the authors propose a systematic strategy to address the challenges of EHR data size imbalance and label imbalance in FL. Their focus is on providing COVID-19 vaccine side effect prediction. In another COVID-19-related work[71], the authors predict mortality for COVID-19 patients using the EHR data from five hospitals. In the context of predictive models in FL, a federated optimization scheme is proposed to predict whether patients with heart-related diseases will be hospitalized within a target year using EHR data. To enhance the security and privacy of EHR data in FL, a more secure system is proposed in[72] to store the data, ensuring confidentiality while enabling collaborative analysis.

IoT data

With the increasing capabilities of internet of things (IoT) devices, health-related data collection has expanded, including periodic heart rate and blood oxygen measurements. These data can be utilized for body condition monitoring, fitness tracking, and elderly care[73]. Since the collected data is often distributed across multiple devices, FL enables the utilization of this data to train machine learning models without the need for data to leave the devices.

In one existing research work[74], a deep FL (DFL) framework is proposed for healthcare data monitoring and analysis in an IoT setting. The framework is tested on a skin disease detection task, and experimental results are provided. In another work[75], an in-home health monitoring system is proposed based on personalized FL. FL frameworks are also proposed to identify individuals' movements using data collected from wearable sensors[76]. Furthermore, several privacy-preserving FL IoT-related healthcare research works[77-79] address privacy and security concerns in IoT-based healthcare by combining FL with blockchain technology[80].

FEDERATED LEARNING FOR RARE DISEASE DETECTION

State-of-the-art work

The issue of imbalance in rare disease detection becomes even more challenging in the FL setting. In reality, each data holder may possess extremely limited patient data related to rare diseases. The distribution of rare disease data can be influenced by factors such as demographics and geographic information. For instance, larger hospitals may have more extensive rare disease data compared to smaller hospitals.

To the best of our knowledge, till this paper, there are limited research works specifically focusing on rare disease detection in FL. In[81], the authors discuss the challenges of EHR data heterogeneity and class imbalance in rare diseases. They propose an FL framework to mitigate the bias caused by imbalanced training data of rare diseases. Specifically, they propose to alleviate the attribute and class biases of the rare disease data by calibrating the feature extractor and the classifier of the models participating in the FL paradigm. In[82], an FL framework is presented for detecting glioblastoma sub-compartment boundaries using data from 6,314 glioblastoma patients across 71 geographically distinct sites spanning six continents. The authors provide detailed descriptions, observations, discussions, and supplementary information regarding their method. Their work shows the utility of federated learning at such scale and task complexity with multiple clients' collaboration without sharing sensitive data. In[83], the authors focus on the inaccuracy tasks and participation of models with different qualities raised by the date property. They apply meta-leaning in the FL framework to predict the rare disease via a proposed dynamic attention and aggregation mechanism, which boots the performance with respect to accuracy and time consumption. In[84], the authors explore the training of ECG and echocardiogram models for hypertrophic cardiomyopathy detection across different institutions in the FL setting. In particular, they combine both ECG and echocardiogram together to help hypertrophic cardiomyopathy detection and show the generalizability across multiple cohorts. In[85], a weakly supervised FL framework for computational pathology is proposed, addressing tasks such as multi-class classification, binary classification, and survival prediction. As in the real-world setting, requiring all the data to be fully labeled is impractical. In addition, the proposed approach provides the capability for each participant to preserve differential privacy by adding random noise. We summarize the existing related works from the dataset, task, and main technique perspectives in Table 1 and will keep maintaining it in the GitHub repository[86].

Future directions and discussion

In this subsection, we discuss the future possible directions of rare disease research in FL.

Lifelong and online machine learning for rare disease in FL

Indeed, the relationships between patients' features and rare diseases are complex and often hidden. Furthermore, the collection of patients' information in each participant within the FL framework occurs over a long period, representing a continuous process. This continuity presents an opportunity to improve the comprehensiveness and accuracy of rare disease diagnosis results.

Lifelong machine learning[87] and online machine learning[88] offer potential solutions for training machine learning models with continuously updated and accumulated data. In the FL setting, participants receive data continuously over the long term. They can update their local models using the accumulated patients' data. Simultaneously, FL allows for further updates through global model aggregation on the server side, followed by distributing the updated models to the client side. This mechanism facilitates the capture and exchange of the latest information on patients with rare diseases. Consequently, it enhances the effectiveness and timeliness of rare disease diagnosis within the FL framework. The integration of lifelong machine learning, online machine learning, and FL leverages the continuous and evolving nature of patient data to improve the accuracy and relevance of rare disease diagnosis.

Multi-modality rare disease detection in FL

As medical technology advances, the ability to collect data from different modalities has increased, providing better support for rare disease detection. However, research institutes, hospitals, and other data holders often possess different modalities of data, such as image data, EHR data, and IoT data. Furthermore, due to the unique properties of rare diseases, there may exist unrevealed and hidden relationships between different features extracted from patients' data.

However, it is unrealistic to assume that each data holder possesses or has the capability to process all modalities of patients' data. Therefore, there is a growing need to develop FL frameworks that can handle different modalities of medical data across various institutions. This framework should aim to maximize the utilization of available data and capture the hidden relationships to effectively support rare disease detection. By leveraging FL, institutions can collaborate and collectively train machine learning models without the need to share sensitive data. The FL framework enables the aggregation of knowledge from different modalities while preserving data privacy and security. This approach allows for a more comprehensive and holistic understanding of rare diseases by integrating diverse data sources. The development of an FL framework that supports multiple modalities of medical data across institutions holds great promise for advancing rare disease detection and improving patient outcomes.

Multi-task rare disease diagnosis in FL

Disease detection is a complex process that often requires collaboration between multiple research resources and knowledge from different domains. This collaboration can involve various departments within the same institution or across different institutions. Each party involved may be responsible for one or more specific tasks, such as analyzing lab results, segmenting X-ray scans, or classifying MRI images. In such scenarios, each participant in FL can utilize their local data to train machine learning models and contribute to

Table 1. Comparison of existing selected related work in rare disease detection with federated learning

Research work	Dataset	Task	Main techniques
Federated learning with imbalanced and agglomerated data distribution for medical image classification[81]	1. Real multi-source dermoscopic image datasets 2. Intracranial hemorrhage classification 3. Skin Lesion Classification	Classification	Feature learning, classifier learning
Federated learning enables big data for rare cancer boundary detection[82]	1. Private data 2. International brain tumor segmentation (BraTS) 2020 challenge	Boundary detection	Data processing, low-resource model running design
DFML: Dynamic federated meta-learning for rare disease prediction[83]	1. Arrhythmia[83] 2. Noninvasive fetal ECG: the physionet/computing in cardiology challenge 2013	Prediction	Meta-learning
Multinational federated learning approach to train ECG and echocardiogram models for hypertrophic cardiomyopathy detection[84]	Private data	Detection	Model generalizability
Federated learning for computational pathology on gigapixel whole slide images[85]	1. Date from the cancer genome atlas 2. Private data	Prediction	Weakly-supervised Learning, differential privacy

collective knowledge without compromising data privacy.

From a global perspective, the server in the FL framework can aggregate the contributions from each participant and generate a comprehensive diagnosis for the targeted rare disease based on the collected information. This aggregation process allows for a holistic view of the disease and leverages the expertise of multiple participants. From the perspective of each participant, the FL framework offers several benefits. Participants can perform their tasks locally, leveraging their own data, expertise, and resources. The FL framework ensures data privacy and security by allowing participants to keep their data within their own environments without the need for direct data sharing. Additionally, participants can benefit from the collective knowledge and insights gained through the collaboration with other participants, enhancing the accuracy and effectiveness of their individual tasks. Overall, an FL framework facilitates efficient collaboration among participants, enabling comprehensive disease diagnosis while preserving data privacy and benefiting each participating party.

Rare disease in heterogenous FL

Rare disease detection in the medical and FL domains presents unique challenges due to its high level of heterogeneity. At the patient level, individuals with the same rare disease can exhibit diverse characteristics, including age, gender, and medical history. Furthermore, at the institutional level, each participant may possess different types and sizes of datasets, with the occurrence and frequency of specific rare disease cases varying based on geographic location and population density[89]. In addition, the popularity density will also have effects on the cases that each institute is able to collect.

Addressing the heterogeneity challenge in rare disease detection requires the design of an FL framework that can effectively handle these variations. The framework should consider the diverse characteristics of patients and their associated data, as well as the differences in rare disease cases and dataset sizes across participating institutions. Strategies need to be developed to accommodate modality missing, task diversity, and variations in data availability to ensure accurate and reliable detection and diagnosis performance. Overcoming heterogeneity in rare disease detection within an FL setting is crucial for leveraging the collective knowledge and resources of participants, ultimately leading to improved outcomes for patients with rare diseases.

Large model-enhanced rare disease diagnosis in FL

In recent years, large models have garnered significant attention and achieved remarkable results in various domains, attracting both industry and academia. These models exhibit the capability to handle diverse types of data and perform complex tasks such as recognition, summarization, generation, question-answering, and even basic communication with humans. Their success can be attributed to the wealth of knowledge they acquire from extensive training data. However, when it comes to the medical field, particularly in the context of rare disease diagnosis, large models often lack specific domain knowledge and access to distributed data held by multiple entities.

The sensitive nature of medical data, coupled with strict data privacy regulations, imposes limitations on directly feeding medical data into large models. Nevertheless, FL presents a promising solution by allowing each local client to leverage the powerful capabilities of large models while preserving data privacy. FL enables collaborative learning across multiple institutions or data holders, where each participant can utilize its local data to contribute to the training process without sharing the actual data. This way, large models can be effectively applied to medical domains, including rare disease diagnosis, by leveraging the expertise and insights gained from clinical notes, EHR data, medical images, and other relevant information. FL bridges the gap between the power of large models and the privacy requirements of medical data, enabling the development of robust and accurate rare disease diagnosis systems.

Human-involved rare disease diagnosis in FL

In the context of rare disease diagnosis, the involvement of human experts, such as doctors, researchers, and domain specialists, is crucial due to the complexity and uniqueness of rare diseases. Considering two practical scenarios, where participants are either hospitals with medical experts or entities with only data and machine learning capabilities, two research directions can be explored to cater to these scenarios.

In the first scenario, where participants have both data and medical experts, it is important to design a mechanism that facilitates the aggregation of comprehensive diagnosis from both machine learning models and human experts in a privacy-preserving manner. One possible approach is to leverage large language models on the server side to aggregate and extract accurate information without directly exchanging sensitive patient data. The server can utilize these models to analyze the local diagnoses provided by each participant, extract relevant insights, and generate feedback to refine and improve the local diagnoses. This collaborative approach ensures that valuable expertise from human professionals is combined with the power of machine learning models, leading to more accurate and comprehensive rare disease diagnosis within each participating institution. In the second scenario, where participants lack local medical experts but possess data and machine learning capabilities, the focus is on leveraging FL mechanisms to support diagnosis by providing additional information and insights. Local models can collaborate through FL to generate diagnostic support, such as uncertainty scores and interpretations, which can be shared with human experts on the server side. This collaboration helps bridge the expertise gap by empowering human professionals with valuable insights from machine learning models. By incorporating the information derived from FL-based models, experts can make more informed decisions and enhance the accuracy and efficiency of rare disease diagnosis.

Both research directions aim to capitalize on the strengths of machine learning models and human expertise in rare disease diagnosis. By carefully designing mechanisms for information aggregation, feedback generation, and collaborative decision-making, FL can facilitate a synergistic relationship between machines and humans, leading to improved diagnosis outcomes while ensuring the privacy and security of sensitive patient data.

PUBLIC DATASETS FOR RARE DISEASE DETECTION

In this section, we will share several existing public rare disease datasets.

Orphanet

Orphanet[90] was established by French National Institute for Health and Medical Research (INSERM) and has been a European endeavor since 2000. It has abundant rare disease resources to support rare disease-related research, diagnosis, treatment, and patient care. Statistically, there are 6,172 rare diseases, 5,835 genes, 8,238 expert centers, and 45,734 diagnostic tests.

OMIM

Online mendelian inheritance in man[91] (OMIM) is a compendium of human genes and genetic phenotypes, which is open access. The database was initially created in the early 1960s and evolved into its online version, OMIM, in 1985, becoming available on the Internet in 1987. It provides online searching functions, basic statistics, and downloads via registration or API access.

MESSIDOR

MESSIDOR stands for Methods to Evaluate Segmentation and Indexing Techniques in the field of Retinal Ophthalmology (in French), which is for computer-assisted diagnoses of diabetic retinopathy. There are 1,200 images from three ophthalmologic departments using a color video 3CCD camera. For each image, there are two diagnoses from the medical experts: retinopathy grade (0,1,2,3) and risk of macular edema (0,1,2). Till this paper, this dataset was last updated on 31 August, 2016.

Others

There are several existing research works where authors combine, organize, link, and extract the existing databases to create easy-to-use datasets. In[92], the authors collected 4,166 rare monogenic diseases and linked them to 3,163 causative genes, which utilizes the information from OMIM, PubMed, Wikipedia, whonamedit.com, and Google Scholar. In the paper, they described the data collection, technical validation, and usage notes.

CONCLUSION

This survey encompasses a comprehensive summary of prevailing AI techniques employed in the detection of rare diseases, alongside an examination of federated learning methodologies within the domain of healthcare informatics. Moreover, we provide an overview of cutting-edge advancements utilizing federated learning for the purpose of rare disease detection, and explore potential topics for future research. To facilitate the validation of newly developed federated learning models, we also compile a selection of relevant datasets. The application of federated learning techniques to the detection of rare diseases not only offers a practical solution but also holds profound significance. We ardently hope that this survey will inspire an increased number of researchers to dedicate their efforts to this vital area of investigation.

DECLARATIONS

Acknowledgments

The authors would like to thank the preliminary work of colleagues from the Data Science Lab at Pennsylvania State University.

Authors' contributions

Made substantial contributions to this research work and manuscript writing: Wang J, Ma F

Availability of data and materials

Not applicable.

Financial support and sponsorship

None.

Conflicts of interest

Both authors declared that there are no conflicts of interest.

Ethical approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Copyright

REFERENCES

1. Rare diseases at FDA. Available from: https://www.fda.gov/patients/rare-diseases-fda [Last accessed on 9 Oct 2023].
2. EU research on rare diseases. Available from: https://research-and-innovation.ec.europa.eu/research-area/health/rare-diseases_en [Last accessed on 9 Oct 2023].
3. What is a rare disease? Available from: https://www.eurordis.org/information-support/what-is-a-rare-disease/ [Last accessed on 9 Oct 2023].
4. Public health challenges of rare diseases. Available from: https://rarediseases.info.nih.gov/ [Last accessed on 9 Oct 2023].
5. Siegel RL, Miller KD, Wagle NS, Jemal A. Cancer statistics, 2023. *CA Cancer J Clin* 2023;73:17-48. DOI PubMed
6. Walter FM, Mills K, Mendonça SC, et al. Symptoms and patient factors associated with diagnostic intervals for pancreatic cancer (SYMPTOM pancreatic study): a prospective cohort study. *Lancet Gastroenterol Hepatol* 2016;1:298-306. DOI PubMed PMC
7. Xing H, Wang J, Wang Y, et al. Diagnostic value of CA 19-9 and carcinoembryonic antigen for pancreatic cancer: a meta-analysis. *Gastroenterol Res Pract* 2018;2018:8704751. DOI PubMed PMC
8. Luo G, Jin K, Deng S, et al. Roles of CA19-9 in pancreatic cancer: biomarker, predictor and promoter. *Biochim Biophys Acta Rev Cancer* 2021;1875:188409. DOI
9. Liu Y, Li L, An S, Helmholz P, Palmer R, Baynam G. 3D face reconstruction with mobile phone cameras for rare disease diagnosis. In: Aziz H, Corrêa D, French T, editors. AI 2022: advances in artificial intelligence. Cham: Springer International Publishing; 2022. pp. 544-56. DOI
10. Lam C, Yu C, Huang L, Rubin D. Retinal lesion detection with deep learning using image patches. *Invest Ophthalmol Vis Sci* 2018;59:590-6. DOI PubMed PMC
11. Sharma S, Malhotra D. A systematic study of intelligent face scanning in rare disease detection. In. Proceedings of 2nd International Conference on Artificial Intelligence: Advances and Applications: ICAIAA 2021: Springer; 2022, p. 451-62. DOI
12. Soni H, Vyas A, Singh U. Identify rare disease patients from electronic health records through machine learning approach. In. 2018 International Conference on Inventive Research in Computing Applications (ICIRCA): IEEE; 2018, p. 1390-5. DOI
13. Garcelon N, Burgun A, Salomon R, Neuraz A. Electronic health records for the diagnosis of rare diseases. *Kidney Int* 2020;97:676-86. DOI PubMed
14. Dong H, Suárez-Paniagua V, Zhang H, et al. Ontology-driven and weakly supervised rare disease identification from clinical notes. *BMC Med Inform Decis Mak* 2023;23:86. DOI PubMed PMC
15. Cheng Y, Liu Y, Chen T, Yang Q. Federated learning for privacy-preserving AI. *Commun ACM* 2020;63:33-6. DOI
16. Li T, Sahu AK, Talwalkar A, Smith V. Federated learning: challenges, methods, and future directions. *IEEE Signal Proc Mag* 2020;37:50-60. DOI
17. Kariampuzha WZ, Alyea G, Qu S, et al. Precision information extraction for rare disease epidemiology at scale. *J Transl Med* 2023;21:157. DOI PubMed PMC
18. Regulation (EC) No 141/2000 of the European Parliament and of the Council of 16 December 1999 on orphan medicinal products. *Official Journal L* 2000;18:1-5. Available from: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A32000R0141 [Last accessed on 9 Oct 2023].
19. Derayeh S, Kazemi A, Rabiei R, Hosseini A, Moghaddasi H. National information system for rare diseases with an approach to data architecture: a systematic review. *Intractable Rare Dis Res* 2018;7:156-63. DOI PubMed PMC

20. Stoller JK. The challenge of rare diseases. *Chest* 2018;153:1309-14. DOI PubMed
21. Cui L, Biswal S, Glass LM, Lever G, Sun J, Xiao C. CONAN: complementary pattern augmentation for rare disease detection. *AAAI* 2020;34:614-21. DOI
22. Li W, Wang Y, Cai Y, Arnold C, Zhao E, Yuan Y. Semi-supervised rare disease detection using generative adversarial network. *arXiv* 2018:181200547. DOI
23. Yu K, Wang Y, Cai Y, et al. Rare disease detection by sequence modeling with generative adversarial networks. *arXiv* 2019:190701022. DOI
24. Yu K, Wang Y, Cai Y. Modelling patient sequences for rare disease detection with semi-supervised generative adversarial nets. In: Lemaire V, Malinowski S, Bagnall A, Bondu A, Guyet T, Tavenard R, editors. Advanced analytics and learning on temporal data. Cham: Springer International Publishing; 2020. pp. 141-50. DOI
25. Arvaniti E, Claassen M. Sensitive detection of rare disease-associated cell subsets via representation learning. *Nat Commun* 2017;8:14825. DOI PubMed PMC
26. Jia X, Xiong Y, Zhang J, Zhang Y, Zhu Y. Few-shot radiology report generation for rare diseases. In. 2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM): IEEE; 2020, p. 601-8. DOI
27. Besson R, Pennec EL, Allassonniere S, et al. A model-based reinforcement learning approach for a rare disease diagnostic task. *arXiv* 2018:181110112. DOI
28. Segura-Bedmar I, Camino-Perdones D, Guerrero-Aspizua S. Exploring deep learning methods for recognizing rare diseases and their clinical manifestations from texts. *BMC Bioinformatics* 2022;23:263. DOI PubMed PMC
29. Prakash P, Chilukuri S, Ranade N, Viswanathan S. RareBERT: transformer architecture for rare disease patient identification using administrative claims. *AAAI* 2021;35:453-60. DOI
30. Li X, Yu L, Jin Y, Fu C, Xing L, Heng P. Difficulty-aware meta-learning for rare disease diagnosis. In: Martel AL, Abolmaesumi P, Stoyanov D, Mateus D, Zuluaga MA, Zhou SK, Racoceanu D, Joskowicz L, editors. Medical image computing and computer assisted intervention-MICCAI 2020. Cham: Springer International Publishing; 2020. pp. 357-66. DOI
31. Lopes RR, Bleijendaal H, Ramos LA, et al. Improving electrocardiogram-based detection of rare genetic heart disease using transfer learning: an application to phospholamban p.Arg14del mutation carriers. *Comput Biol Med* 2021;131:104262. DOI
32. Taroni JN, Grayson PC, Hu Q, et al. MultiPLIER: a transfer learning framework for transcriptomics reveals systemic features of rare disease. *Cell Syst* 2019;8:380-94.e4. DOI PubMed PMC
33. Li J, He Z, Zhang M, et al. Estimating rare disease incidences with large-scale internet search data: development and evaluation of a two-step machine learning method. *JMIR Infodemiology* 2023;3:e42721. DOI PMC
34. Spiga O, Cicaloni V, Fiorini C, et al. Machine learning application for development of a data-driven predictive model able to investigate quality of life scores in a rare disease. *Orphanet J Rare Dis* 2020;15:46. DOI PubMed PMC
35. Lee J, Liu C, Kim J, et al. Deep learning for rare disease: a scoping review. *J Biomed Inform* 2022;135:104227. DOI
36. Schaefer J, Lehne M, Schepers J, Prasser F, Thun S. The use of machine learning in rare diseases: a scoping review. *Orphanet J Rare Dis* 2020;15:145. DOI PubMed PMC
37. Banerjee J, Taroni JN, Allaway RJ, Prasad DV, Guinney J, Greene C. Machine learning in rare disease. *Nat Methods* 2023;20:803-14. DOI PubMed
38. Decherchi S, Pedrini E, Mordenti M, Cavalli A, Sangiorgi L. Opportunities and challenges for machine learning in rare diseases. *Front Med* 2021;8:747612. DOI PubMed PMC
39. Brasil S, Pascoal C, Francisco R, Dos Reis Ferreira V, Videira PA, Valadão AG. Artificial intelligence (AI) in rare diseases: is the future brighter? *Genes* 2019;10:978. DOI PubMed PMC
40. Groft SC, Posada M, Taruscio D. Progress, challenges and global approaches to rare diseases. *Acta Paediatr* 2021;110:2711-6. DOI PubMed
41. Brasil S, Neves CJ, Rijoff T, et al. Artificial intelligence in epigenetic studies: shedding light on rare diseases. *Front Mol Biosci* 2021;8:648012. DOI PubMed PMC
42. Hurvitz N, Azmanov H, Kesler A, Ilan Y. Establishing a second-generation artificial intelligence-based system for improving diagnosis, treatment, and monitoring of patients with rare diseases. *Eur J Hum Genet* 2021;29:1485-90. DOI PubMed PMC
43. Skweres-Kuchta M, Czerska I, Szaruga E. Literature review on health emigration in rare diseases-a machine learning perspective. *Int J Environ Res Public Health* 2023;20:2483. DOI PubMed PMC
44. Visibelli A, Roncaglia B, Spiga O, Santucci A. The impact of artificial intelligence in the odyssey of rare diseases. *Biomedicines* 2023;11:887. DOI PubMed PMC
45. Roman-Naranjo P, Parra-Perez AM, Lopez-Escamez JA. A systematic review on machine learning approaches in the diagnosis of rare genetic diseases. *medRxiv* 2023:23285203. DOI
46. Hasani N, Farhadi F, Morris MA, et al. Artificial intelligence in medical imaging and its impact on the rare disease community: threats, challenges and opportunities. *PET Clin* 2022;17:13-29. DOI PubMed PMC
47. McMahan B, Moore E, Ramage D, Hampson S, y Arcas BA. Communication-efficient learning of deep networks from decentralized data. *PMLR* 2017;54:1273-82. DOI
48. Wang J, Zeng S, Long Z, Wang Y, Xiao H, Ma F. Knowledge-enhanced semi-supervised federated learning for aggregating heterogeneous lightweight clients in IoT. In: Shekhar S, Zhou Z, Chiang Y, Stiglic G, editors. Proceedings of the 2023 SIAM International Conference on Data Mining (SDM). Philadelphia: Society for Industrial and Applied Mathematics; 2023. pp. 496-504.

DOI

49. Tan Y, Long G, Liu L, et al. FedProto: federated prototype learning across heterogeneous clients. *AAAI* 2022;36:8432-40. DOI
50. Jere MS, Farnan T, Koushanfar F. A taxonomy of attacks on federated learning. *IEEE Secur Privacy* 2021;19:20-8. DOI
51. Solanki S, Kanaparthy S, Damle S, Gujar S. Differentially private federated combinatorial bandits with constraints. In: Amini M, Canu S, Fischer A, Guns T, Kralj Novak P, Tsoumakas G, editors. Machine learning and knowledge discovery in databases. Cham: Springer Nature Switzerland; 2023. pp. 620-37. DOI
52. Lycklama H, Burkhalter L, Viand A, Küchler N, Hithnawi A. RoFL: robustness of secure federated learning. In. 2023 IEEE Symposium on Security and Privacy (SP): IEEE Computer Society; 2023, p. 453-76. DOI
53. Lyu L, Xu X, Wang Q, Yu H. Collaborative Fairness in federated learning. In: Yang Q, Fan L, Yu H, editors. Federated learning. Cham: Springer International Publishing; 2020. pp. 189-204. DOI
54. Wang J, Qian C, Cui S, Glass L, Ma F. Towards federated COVID-19 vaccine side effect prediction. In: Amini M, Canu S, Fischer A, Guns T, Kralj Novak P, Tsoumakas G, editors. Machine learning and knowledge discovery in databases. Cham: Springer Nature Switzerland; 2023. pp. 437-52. DOI
55. Liu Y, Huang A, Luo Y, et al. FedVision: an online visual object detection platform powered by federated learning. *AAAI* 2020;34:13172-9. DOI
56. Huang A, Liu Y, Chen T, et al. StarFL: hybrid federated learning architecture for smart urban computing. *ACM Trans Intell Syst Technol* 2021;12:1-23. DOI
57. Long G, Tan Y, Jiang J, Zhang C. Federated learning for open banking. In: Yang Q, Fan L, Yu H, editors. Federated learning. Cham: Springer International Publishing; 2020. pp. 240-54. DOI
58. IBM federated learning. Available from: https://www.ibm.com/docs/en/cloud-paks/cp-data/4.7.x?topic=models-federated-learning [Last accessed on 9 Oct 2023].
59. Federated learning on Google Cloud. Available from: https://cloud.google.com/architecture/federated-learning-google-cloud [Last accessed on 9 Oct 2023].
60. TensorFlow federated: machine learning on decentralized data. Available from: https://www.tensorflow.org/federated [Last accessed on 9 Oct 2023].
61. Dang TK, Lan X, Weng J, Feng M. Federated learning for electronic health records. *ACM Trans Intell Syst Technol* 2022;13:1-17. DOI
62. Xu J, Glicksberg BS, Su C, Walker P, Bian J, Wang F. Federated learning for healthcare informatics. *J Healthc Inform Res* 2021;5:1-19. DOI PubMed PMC
63. Antunes RS, André da Costa C, Küderle A, Yari IA, Eskofier B. Federated learning for healthcare: systematic review and architecture proposal. *ACM Trans Intell Syst Technol* 2022;13:1-23. DOI
64. Oh W, Nadkarni GN. Federated learning in health care using structured medical data. *Adv Kidney Dis Health* 2023;30:4-16. DOI PubMed PMC
65. Pfitzner B, Steckhan N, Arnrich B. Federated learning in a medical context: a systematic literature review. *ACM Trans Internet Technol* 2021;21:1-31. DOI
66. Rieke N, Hancox J, Li W, et al. The future of digital health with federated learning. *NPJ Digit Med* 2020;3:119. DOI PubMed PMC
67. Reddy K, Gadekallu TR. A comprehensive survey on federated learning techniques for healthcare informatics. *Comput Intell Neurosci* 2023;2023:8393990. DOI PubMed PMC
68. Sheller MJ, Reina GA, Edwards B, Martin J, Bakas S. Multi-institutional deep learning modeling without sharing patient data: a feasibility study on brain tumor segmentation. In: Crimi A, Bakas S, Kuijf H, Keyvan F, Reyes M, van Walsum T, editors. Brainlesion: glioma, multiple sclerosis, stroke and traumatic brain injuries. Cham: Springer International Publishing; 2019. pp. 92-104. DOI PubMed PMC
69. Feki I, Ammar S, Kessentini Y, Muhammad K. Federated learning for COVID-19 screening from Chest X-ray images. *Appl Soft Comput* 2021;106:107330. DOI PubMed PMC
70. Linardos A, Kushibar K, Walsh S, Gkontra P, Lekadir K. Federated learning for multi-center imaging diagnostics: a simulation study in cardiovascular disease. *Sci Rep* 2022;12:3551. DOI PubMed PMC
71. Vaid A, Jaladanki SK, Xu J, et al. Federated learning of electronic health records to improve mortality prediction in hospitalized patients with COVID-19: machine learning approach. *JMIR Med Inform* 2021;9:e24207. DOI PubMed PMC
72. Salim MM, Park JH. Federated learning-based secure electronic health record sharing scheme in medical informatics. *IEEE J Biomed Health Inform* 2023;27:617-24. DOI PubMed
73. Islam SM, Kwak D, Kabir H, Hossain M, Kwak KS. The internet of things for health care: a comprehensive survey. *IEEE Access* 2015;3:678-708. DOI
74. Elayan H, Aloqaily M, Guizani M. Sustainability of healthcare data analysis IoT-based systems using deep federated learning. *IEEE Internet Things J* 2022;9:7338-46. DOI
75. Wu Q, Chen X, Zhou Z, Zhang J. FedHome: cloud-edge based personalized federated learning for in-home health monitoring. *IEEE Trans on Mobile Comput* 2022;21:2818-32. DOI
76. Arikumar KS, Prathiba SB, Alazab M, et al. FL-PMI: federated learning-based person movement identification through wearable devices in smart healthcare systems. *Sensors* 2022;22:1377. DOI PubMed PMC
77. Li J, Meng Y, Ma L, et al. A federated learning based privacy-preserving smart healthcare system. *IEEE T Ind Inform* 2021;18:2021-

31. DOI

78. Kurniawan H, Mambo M. Homomorphic encryption-based federated privacy preservation for deep active learning. *Entropy* 2022;24:1545. DOI PubMed PMC
79. Alam T, Gupta R. Federated learning and its role in the privacy preservation of iot devices. *Future Internet* 2022;14:246. DOI
80. Singh S, Rathore S, Alfarraj O, Tolba A, Yoon B. A framework for privacy-preservation of IoT healthcare data using federated learning and blockchain technology. *Future Gener Comp Sy* 2022;129:380-8. DOI
81. Wu N, Yu L, Yang X, Cheng K-T, Yan Z. FedIIC: towards robust federated learning for class-imbalanced medical image classification. *arXiv* 2022:220613803. DOI
82. Pati S, Baid U, Edwards B, et al. Federated learning enables big data for rare cancer boundary detection. *Nat Commun* 2022;13:7346. DOI
83. Chen B, Chen T, Zeng X, et al. DFML: dynamic federated meta-learning for rare disease prediction. *IEEE/ACM Trans Comput Biol Bioinform* 2023:1-11. DOI
84. Goto S, Solanki D, John JE, et al. Multinational federated learning approach to train ecg and echocardiogram models for hypertrophic cardiomyopathy detection. *Circulation* 2022;146:755-69. DOI PubMed PMC
85. Lu MY, Chen RJ, Kong D, et al. Federated learning for computational pathology on gigapixel whole slide images. *Med Image Anal* 2022;76:102298. DOI PubMed PMC
86. Wang J. Rare disease detection with federated learning. Available from: https://github.com/JackqqWang/rare_disease_detection_fl/blob/main/main.md [Last accessed on 10 Oct 2023].
87. Thrun S. Lifelong learning algorithms. In: Thrun S, Pratt L, editors. Learning to learn. Boston: Springer US; 1998. pp. 181-209. DOI
88. Fontenla-Romero Ó, Guijarro-Berdiñas B, Martinez-Rego D, Pérez-Sánchez B, Peteiro-Barral D. Online machine learning. In: editor^editors, editor. Efficiency and scalability methods for computational intellect:IGI global;2013.p.27-54. DOI
89. Chen C, Liang J, Ma F, Glass L, Sun J, Xiao C. Unite: uncertainty-based health risk prediction leveraging multi-sourced data. In. Proceedings of the Web Conference 2021; 2021, p. 217-26. DOI
90. The portal for rare diseases and orphan drugs. Available from: https://www.orpha.net/ [Last accessed on 9 Oct 2023].
91. An online catalog of human genes and genetic disorders. Available from: https://www.omim.org/ [Last accessed on 9 Oct 2023].
92. Ehrhart F, Willighagen EL, Kutmon M, van Hoften M, Curfs LMG, Evelo CT. A resource to explore the discovery of rare diseases and their causative genes. *Sci Data* 2021;8:124. DOI PubMed PMC