IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 12 December 2023, accepted 13 January 2024, date of publication 23 January 2024, date of current version 1 February 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3357514

RESEARCH ARTICLE

FedCure: A Heterogeneity-Aware Personalized Federated Learning Framework for Intelligent Healthcare Applications in IoMT Environments

SACHIN D. N[1], ANNAPPA B[1], (Senior Member, IEEE), SAUMYA HEGDE[1], (Senior Member, IEEE), CHUNDURU SRI ABHIJIT[2], AND SATEESH AMBESANGE[1]

[1]National Institute of Technology Karnataka, Surathkal, Mangalore 575025, India
[2]Vellore Institute of Technology, Chennai 600127, India

Corresponding author: Sachin D. N (sachindn.207cs004@nitk.edu.in)

ABSTRACT The advent of the Internet of Medical Things (IoMT) devices has led to a healthcare revolution, introducing a new era of smart applications driven by Artificial Intelligence (AI). These advanced technologies have greatly influenced the healthcare industry and have played a crucial role in enhancing the quality of life globally. Federated Learning (FL) has become popular as a technique to create models that can be shared universally using the vast datasets collected from IoMT devices while maintaining data privacy. However, the complex variations in IoMT environments, including diverse devices, data characteristics, and model complexities, create challenges for the straightforward application of traditional FL methods. Consequently, it is not well-suited for deployment in such contexts. This paper introduces FedCure, a personalized FL framework tailored for intelligent IoMT-based healthcare applications operating within a cloud-edge architecture. FedCure is adept at addressing the challenges within IoMT environments by employing personalized FL techniques that can effectively mitigate the impact of heterogeneity. Furthermore, the integration of edge computing technology enhances processing speed and minimizes latency in intelligent IoMT applications. Lastly, this research showcases several case studies encompassing IoMT-based applications, such as Eye Retinopathy Detection, Diabetes Monitoring, Maternal Health, Remote Health Monitoring, and Human Activity Recognition. These case studies provide a means to assess the effectiveness of the proposed FedCure framework and showcase exceptional performance with accuracy and minimal communication overhead, especially in addressing the challenges posed by heterogeneity.

INDEX TERMS Federated learning, edge computing, digital healthcare, Internet of Medical Things (IoMT).

I. INTRODUCTION

The Internet of Medical Things (IoMT) is a rapidly evolving field that has captured the interest of researchers in machine learning (ML) and healthcare technology for more than a decade [1]. IoMT is a transformative paradigm within healthcare, leveraging interconnected medical devices to enhance patient care, streamline medical processes, and drive innovation in healthcare solutions [2]. This convergence of technology and medicine has ushered in significant advancements in the digital healthcare landscape [3]. IoMT's impact within the healthcare sector is profound, driven by many interconnected devices and systems that revolutionize patient care and medical processes [4]. These functions encompass data collection, real-time monitoring, diagnostics, treatment, and disease management. IoMT has empowered individuals by granting them greater access to medical data, fostering patient empowerment and proactive healthcare management [3]. One pivotal aspect of IoMT's transformative power lies in medical data management and analysis.

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

The vast volume of healthcare data generated by IoMT devices, including remote monitoring equipment, wearable fitness trackers, and intelligent medical devices, presents a wealth of valuable information [5]. ML algorithms are at the core of the Internet of Medical Things (IoMT) and have been instrumental in analyzing and interpreting the massive amounts of data generated by it. One of the key areas where IoMT-driven ML models have been successful is in medical image segmentation. These models can accurately identify and outline structures within medical images, which is useful in diagnosing and planning treatments for medical conditions [5]. These models also demonstrate impressive capabilities in disease classification, facilitating the rapid and accurate categorization of ailments based on symptom profiles and patient data. Additionally, IoMT aids in disease detection, enabling early diagnosis and intervention, ultimately enhancing patient outcomes [4]. There are significant challenges when it comes to using AI to train patient data, such as data privacy and distributed data learning. In order to train models, patient data needs to be shared, which raises concerns about privacy. Regulations like GDPR [6] and HIPAA [7] ensure secure handling of data, but can also lead to fragmentation of data, making it difficult to share and train AI models. Another challenge lies in the distributed nature of IoMT data. It originates from diverse sensors and devices, typically not centralized. Consequently, data from various hospitals and servers must be trained in a distributed manner. These data sources often exhibit distinct patterns or distributions, further complicating the learning process [8]. These challenges form the backdrop against which this research seeks to contribute.

To address this challenge, Federated Learning (FL) offers an evolved solution [9]. FL presents a sophisticated mechanism for the collaborative training of a high-caliber shared model. This collaborative approach involves aggregating locally computed updates contributed by IoMT devices. A major benefit of this methodology is its capability to separate the model training process from direct access to training data. Essentially, FL empowers the creation of a dependable global model while upholding the utmost privacy of user data. However, it is crucial to acknowledge that the intricate nature of IoMT environments poses significant impediments to the seamless integration of FL, rendering its direct implementation in IoMT applications a challenging endeavor. The main challenges using FL in complex IoMT environments are data, device, and model heterogeneity [10]. First, the data heterogeneity arises due to IoMT devices collecting various health data, such as heart rate or blood pulse, resulting in distinct data patterns influenced by individual habits. Second, the devices differ in storage, processing power, and communication abilities, complicating data management. Lastly, diverse devices require unique models to suit their specific applications; for instance, some may only support basic models due to limited resources, leading to communication and performance issues. Traditional FL methods often encounter difficulties addressing these challenges when applied to collaborative learning in IoMT networks. While researchers have put in a lot of effort to deal with the problem of heterogeneity by suggesting the result of a global model that combines knowledge from all devices involved, this approach falls short in preserving the unique information from each device [9], [11], [12]. Consequently, it leads to a decline in performance when making predictions or classifications. Yet, traditional FL methods operate under the assumption of ample data and resources at the edge, a condition often unmet in real-world situations, particularly within IoMT-based healthcare networks. For instance, considering the issue of data heterogeneity, some hospitals have more sensitive patient information, while others have mostly normal data. This discrepancy creates ‘‘data partitions with non-identical distribution’’ [13].

To address these heterogeneity challenges effectively, a promising approach involves implementing personalization at the device, data, and model levels. This strategy helps alleviate disparities and achieve high-quality personalized models tailored to each device. Recently, there’s been growing interest in Personalized FL. Researchers have suggested these frameworks to address the issue of data differences [14], [15], [16]. However, there haven’t been many experiments on healthcare applications with IoMT networks with diverse data, so there’s room for more research in this area.

This paper delved into examining emerging personalized FL approaches. It proposed FedCure, a cloud-edge-based framework for personalized FL offering promising solutions to address various heterogeneity challenges within complex IoMT environments. These techniques hold significant potential for enabling the development of intelligent IoMT applications. Edge computing allows IoMT devices to delegate computationally intensive learning tasks to the edge, resulting in fast processing and minimal latency despite device heterogeneity. This framework adopted different personalized FL approaches at the device level to deploy customized models tailored to specific application needs. The suggested framework underwent evaluation through various case studies involving IoMT applications, such as Human Activity Recognition, Eye Retinopathy Classification, Fitness Tracking, Diabetic Prediction, and Mental Health Analysis. The outcomes highlight remarkable performance in terms of accuracy with minimal communication overhead.

The main contribution of this paper is outlined below:

- Proposed FedCure: a cloud-edge-based personalized FL framework for Intelligent healthcare applications in complex IoMT environments.
- Data heterogeneity: FedCure addresses challenges stemming from data and model heterogeneity by implementing various personalization techniques at the device level, facilitating the deployment of customized models tailored to specific requirements.
- Device heterogeneity: Additionally, to mitigate issues related to device heterogeneity, the framework incorporates computational offloading techniques using

edge computing, ensuring efficient processing and optimization.
- Performance & Evaluation: Testing the proposed framework on various IoMT-based healthcare application use cases, it outperformed in handling heterogeneity issues in complex IoMT environments.

The paper adheres to a specific structure, as outlined below. In the upcoming section II, the paper delves into the primary challenges of implementing FL in IoMT environments. To tackle these challenges, the proposed work introduces a personalized FL framework based on a cloud-edge architecture, as detailed in Section III. Various emerging personalization solutions are explored within this section. Subsequently, in Section V, the paper presents a study encompassing different case scenarios and evaluates the efficacy of personalized FL methods through practical case studies. Finally, the concluding remarks on the proposed work are provided in Section VI.

II. BACKGROUND AND MOTIVATION

In this section, a detailed exploration will be undertaken to examine the various applications of FL-based healthcare technologies to improve patient outcomes. Additionally, we will address the intricacies of implementing these solutions within complex IoMT environments, where challenges arise due to varying data distributions, learning tasks, communication complexities, and computational issues.

A. FL-IOMT BASED HEALTHCARE

In modern healthcare applications, IoMT systems play an integral role. These systems are multifaceted and designed to perform various functions, including data acquisition and local storage, communication with other devices, data processing, analysis, and global data storage, as elucidated by Gupta et al. [17]. Khowaja et al. [1] contribution in healthcare is underscored by their ability to provide real-time patient monitoring, facilitate remote diagnostics, and ultimately enhance patient outcomes. However, this domain has challenges as IoMT systems grapple with interoperability, security, mobility, standardization, and licensing issues. These challenges have driven research and innovation, yielding various solutions to safeguard sensitive medical data. Prominent among these solutions are cutting-edge technologies, including Blockchain, Access Control models, and Homomorphic Encryption, as outlined by Ahmed et al. [18].

FL emerges as a promising approach to tackle some of these IoMT challenges. FL [9] operates by collaboratively harnessing data from numerous mobile devices worldwide. This collaborative approach holds great potential for enhancing the efficiency and security of IoMT systems while rigorously safeguarding the privacy and integrity of medical data. Nevertheless, the intricacies of IoMT, such as device heterogeneity, can introduce its challenges, including communication bottlenecks, straggler problems, and faulty nodes [18]. Notable IoMT frameworks have been developed to address specific healthcare needs. For instance, MyWear [19] employs smart garments to monitor patient vitals and predict heart failure risk continuously. In contrast, iLog [20] identifies food intake and stress levels through an innovative IoMT solution.

Several frameworks and architectures have been proposed in the realm of healthcare technology. One example is FedHome [21], aimed at in-home health monitoring, utilizing a generative convolutional autoencoder (GCAE). Alzubi et al. [22] introduced a cloud-based and blockchain-enabled FL architecture to preserve the privacy of electronic health records. Additionally, Lu et al. [23] proposed a personalized FL system catering to the distinct data distributions of clients. Despite the notable advancements in FL-based systems within the IoMT context, challenges still impact the effectiveness of healthcare services. These include addressing issues like sparse data, diverse user behavior, and system heterogeneity, which have not been comprehensively explored in existing literature. Moreover, it is essential to highlight that previous research efforts are yet to create a personalized FL-enabled heterogeneous IoMT system tailored specifically for healthcare applications.

B. CHALLENGES OF USING FL IN HETEROGENEOUS IOMT ENVIRONMENTS

IoMT environments require an IoMT-oriented FL framework. The existing FL works are not derived from genuine IoMT devices, which makes it vital to bridge the gap by developing an IoMT-oriented FL framework. The main challenge is that healthcare data comes from various sources, leading to heterogeneity. In a supervised task, when training a model with user data distribution represented as $P_i(x, y)$, x usually signifies the input features or attributes of the data. Meanwhile, y represents the corresponding class labels or target variable. It is crucial to recognize that data collected from different devices may not conform to the same distribution pattern because of varying environments. User data can exhibit differences in several aspects, including the distribution of features, assignment of labels, and shifts in underlying concepts [24]. This complicates the optimization process and increases the risk of straggler clients. In an IoMT-equipped healthcare setting, patients can access diverse data and health parameters. Training AI models on data from a single hospital or clinic can introduce bias. Smaller healthcare clinics may have limited Electronic Health Record (EHR) data, leading to data sparsity. Nonetheless, it is worth noting that FL methods are typically evaluated using compact datasets with restricted characteristics [25], [26].

Using IoMT devices from different vendors with varying hardware, software, and training platforms presents significant challenges to FL. These challenges include problems with data communication due to the heterogeneity of data

characteristics and dimensions, leading to discreteness in local data storage formats [17]. An effective FL framework for IoMT should be able to adjust to these various sources of diversity to enhance performance. In an FL environment, slow or expensive connections, offline devices, and devices with limited computing capacity can become a communication constraint, causing participating devices to drop out due to poor connectivity and energy constraints [27]. Therefore, addressing the communication and computation issues of heterogeneous devices in the FL-based IoMT environment is crucial. FL uses customized models tailored to individual devices, leading to model heterogeneity. Within IoMT, diverse devices aim to construct adaptable models driven by the distinctive requirements of their application environments and limitations in available resources. However, privacy concerns prevent model sharing, leading to different model architectures that cannot be naively aggregated using traditional FL [28]. This can manifest in different ways, such as varying neural network architectures, optimization techniques, or learning rates. Managing heterogeneity is crucial for efficient collaboration and success in diverse environments [26].

AI-based IoMT is a rapidly growing technology used in complex healthcare environments to monitor individuals. However, since this technology involves collecting sensitive patient data, it is necessary to encrypt the data before transmitting it to central servers to ensure privacy. Sending these healthcare data over networks can be inefficient and incur substantial costs. Additionally, training healthcare data solely on individual devices can lead to accuracy challenges [29]. To address these concerns, an FL approach is used, which involves conducting local data training and transmitting only the trained models to central servers. This approach helps to address accuracy and privacy issues. Conventional FL techniques might not be well-suited for devices with limited resources operating in the heterogeneous IoMT settings, marked by diverse data types and distributions among edge devices. FL can be used to extract common knowledge from all devices and create a high-quality global model. Nonetheless, it falls short of capturing individualized information, leading to a decline in inference quality. In the complex landscape of IoMT applications, achieving universal consensus among devices for a shared model is impractical. Each edge device holds valuable and sensitive information crucial in IoMT-based healthcare applications. Challenges in IoMT environments arise from data and device issues, hindering the collaborative development of intelligent applications while ensuring data privacy. Conventional FL approaches often fall short in efficiently addressing these challenges. Our research is driven by the goal of preserving all critical information within complex healthcare settings by involving all IoMT devices in the FL process. Our primary focus is on tackling the data and device heterogeneity present in IoMT networks, a challenge inadequately handled by standard FL methods.

III. PERSONALIZED FL FRAMEWORK FOR HETEROGENEOUS IOMT ENVIRONMENTS

In the complex IoMT network landscape, interconnectivity spans a vast ecosystem. This intricate network seamlessly links cloud servers, various hospital medical research labs, mobile health application servers, and numerous healthcare organizations. Each healthcare entity employs many IoMT devices within this extensive infrastructure, generating diverse heterogeneous data. The difficulty arises because each IoMT device is uniquely configured w.r.t data processing capabilities and capacity, resulting in a highly intricate data environment. The conventional FL approach faces substantial obstacles when dealing with this unprecedented heterogeneity. Device heterogeneity, stemming from differences in data generation and transmission capabilities, introduces a layer of complexity that traditional FL struggles to address. Statistical heterogeneity arises from the varied data distributions in healthcare entities and research labs, further complicating the FL process. Additionally, model heterogeneity, driven by the differing configurations of IoMT devices, adds another layer of intricacy. Consequently, a global FL model may not perform optimally across a multifaceted environment. A compelling solution to address these heterogeneity issues is personalization. By tailoring learning models to the unique characteristics of each IoMT device, cloud server, or healthcare organization, we can effectively manage the complexity of data distributions, device capabilities, and network structures. In this context, personalization becomes a pivotal strategy for enhancing data analysis and processing efficiency and accuracy within the IoMT network. Advanced FL techniques enable personalized models for IoMT devices with resource constraints while facilitating collective knowledge sharing. The need for adaptable and personalized models becomes self-evident within the intricate IoMT ecosystem characterized by diverse devices and complex data landscapes. These personalized models can be fine-tuned to cater to the specific demands of each device, thereby ensuring the optimal utilization of resources and superior performance. By integrating the principles of FL, the network can effectively coordinate these personalized models and promote collaboration among devices, allowing them to pool their knowledge resources while maintaining their unique models. This synergistic approach, combining advanced FL methods with personalized modeling, promises to enable the IoMT network to navigate the challenges of device heterogeneity, statistical disparities, and model variations. It provides the framework for a data-driven collaborative environment where each component can flourish while significantly contributing to the collaborative learning pool. This paper introduces a tailored FL framework designed for intelligent IoMT applications, with the primary goal of comprehensively tackling the challenges posed by heterogeneity.

In the depicted scenario, Figure 1 illustrates a complex IoMT-based healthcare environment structured into three

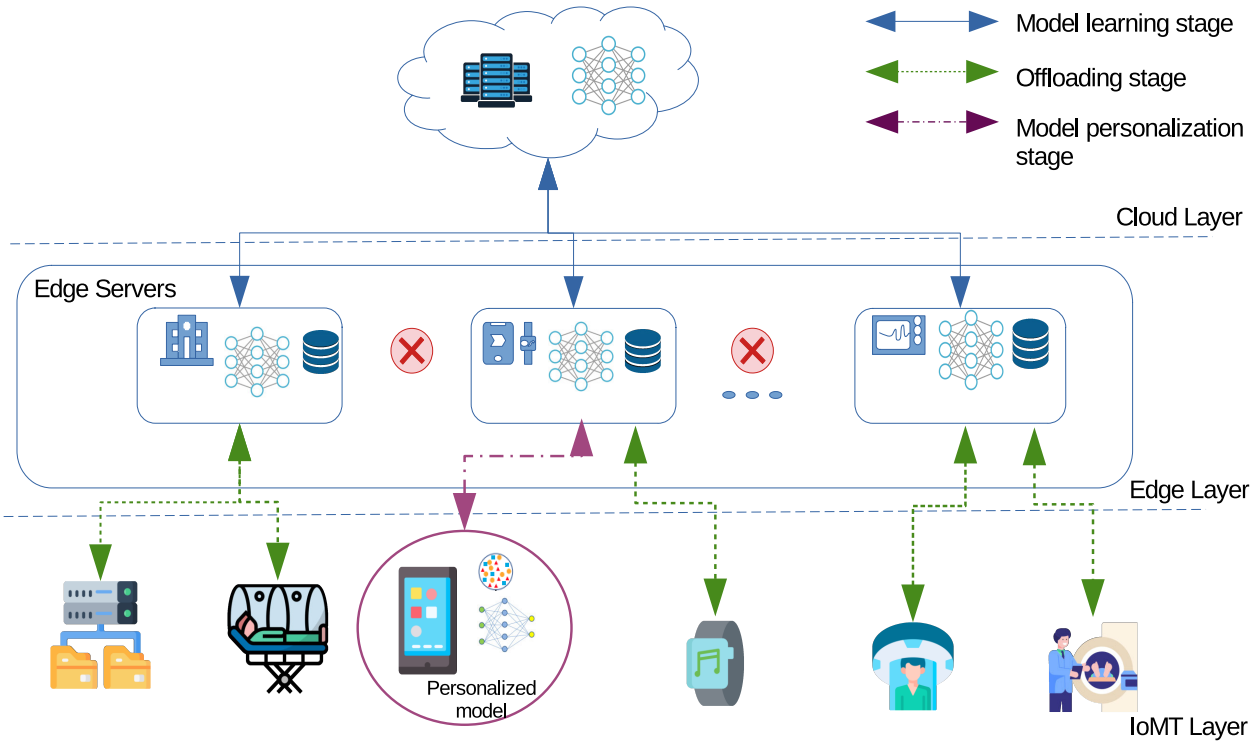

FIGURE 1. FedCure framework.

distinct layers: cloud, edge, and IoMT devices. This framework is designed to facilitate the development of intelligent healthcare applications while maintaining data privacy by transmitting data to a centralized server. In the proposed FedCure framework, a FL server is situated in the cloud layer, is responsible for coordinating the model training process and consolidating model updates from different sources. The second layer, termed the 'edge layer,' is purpose-built to provide data processing capabilities close to the data sources. This layer comprises different healthcare servers, each serving a range of healthcare institutions and organizations, ensuring efficient data processing. Each healthcare edge server is connected to different IoMT devices, which generate various types of distributed data. Finally, the third layer, known as the 'IoMT layer,' encompasses various healthcare devices. These devices are tasked with collecting and transmitting data related to healthcare and medical parameters, making them an indispensable component of the overall ecosystem.

In the proposed intelligent IoMT ecosystem, in order to enable efficient computation and alleviate the computational load on IoMT devices, the architecture allows these devices to offload their computational tasks and data to trusted healthcare edge servers via network connections. This strategy enables IoMT devices to harness the computational capabilities offered by edge servers, guaranteeing the fulfillment of requirements for high processing efficiency and minimal latency in healthcare applications. Additionally, this strategy maintains data privacy and security while enhancing the overall performance of intelligent healthcare systems in order to facilitate collaborative learning by using FL. With this method, edge servers, the distant cloud, and IoMT devices jointly train a global model. This maintains the privacy of sensitive data on individual devices while tailoring the learning model to each device's unique capabilities.

In the proposed framework, collaborative learning occurs in three stages: offloading, learning, and model personalization. These stages solve heterogeneity issues in complex IoMT network-based intelligent applications. In the offloading stage, IoMT devices from hospitals and healthcare organizations collaborate to train edge models in our healthcare ecosystem. Devices with limited computational resources offload their tasks and data to edge servers for collective model training [30]. The global loss function is expressed as:

$$F(w) = \frac{1}{|D|} \sum_{i \in D} f_i(w) \tag{1}$$

where the dataset size is denoted by $|D|$, the global loss is represented by $F(w)$, and the specific loss for each data point i is denoted by $f_i(w)$. Finding the ideal model variables w^* that minimize this global loss is the main goal, and this is defined as:

$$w^* = \arg\min_w F(w) \tag{2}$$

In addition to this objective, the system delay encompasses data offloading and model training delays. Data offloading delay T_{EL} takes in accounts for bandwidth B, transmission power p_k, channel gain g_k, and more. The model delay includes server computation W and CPU frequency e_s, where $W = Ne \cdot C_D$. Here, Ne is the number of epochs, C_D is the number of CPU cycles needed for 1-bit data, and D is the dataset.

During the learning phase, all participants, edge servers, and devices independently conduct updates based on their respective local datasets. Each participant exclusively sends their model parameters to the central server during this process. Following this, the server combines these newly updated parameters to create a global model, which is then shared with the participants for subsequent local updates. Before aggregating these parameters, each participant may undertake one or more training epochs as part of their local update phase, commonly called a 'communication round.' For a specific participant indexed as k, the loss function is represented as follows:

$$F_k(w_k) = \frac{1}{|D_k|} \sum_{i \in D_k} f_i(w_k) \tag{3}$$

In this context, w_k represents the local model parameter specific to client k. Following the principles outlined in the FedAvg algorithm [9], the global model parameter is defined as:

$$w = \frac{1}{D} \sum_{k \in K} D_k w_k \tag{4}$$

This iterative process continues until convergence is achieved. After obtaining a high-quality global model, it is transmitted back to clients for further personalization.

The personalization stage is a way to balance inter-client collaboration and individual performance. Standard FL uses a global model for all clients without personalization. Nevertheless, this method may prove less effective when client data distributions are not uniform. Each client trains its model in a local learning setting, resulting in a fully personalized model for each client θ_k. The objective in this setup is to optimize and fine-tune their model without considering inputs or contributions from other clients.

$$\min_{\theta_1,\ldots,\theta_k \in \mathbb{R}^d} F(\theta) = \frac{1}{|D_k|} \sum_{i \in D_k} f_i(\theta_k) \tag{5}$$

Here, $\theta_k \in \mathbb{R}^d$ signifies the local model parameters unique to client k. Nevertheless, this method may not attain optimal generalization performance. Personalized FL approaches strike a balance between conventional FL and local learning settings. They enable clients to collaborate and share knowledge while delivering personalized outcomes for each client.

Another pressing issue emerges in the complex IoMT network, where device diversity and computational limitations create hurdles. It pertains to the varied data characteristics - differing in distribution, quality, and quantity, and often not adhering to the standard IID (independently and identically distributed) pattern. Non-IID data can also exhibit label and feature imbalances, adding complexity. Furthermore, data on each node follows distinct distributions with varying data points. Managing and utilizing these various data may be difficult due to the possibility of an underlying structure connecting these nodes and their data patterns. To tackle these complexities, this paper introduces the FedCure framework. It uses edge computing to improve the capabilities of individual devices by offloading computations addressing straggling. Additionally, we reduce the communication workload by aggregating local models at the edge server. FedCure's adaptability allows it to seamlessly incorporate various personalized federated methods, enabling the exchange of diverse model information between edge devices and the cloud. These strategies aim to streamline data processing and sharing, mitigating heterogeneity and complexity within the IoMT network. FedCure is ideal for large-scale practical healthcare applications in the IoMT field.

IV. PERSONALIZED FEDERATED LEARNING (PFL) APPROACHES

This section examined and provided detailed insights into various essential personalized FL methods that can be seamlessly combined with the FedCure framework in the context of intelligent IoMT applications.

A. HYPERNETWROKS BASED PFL

Hypernetworks (HN) [31] are neural networks capable of generating weights and architectures for other networks. They find their application in various ML fields. HNs play a vital role in adapting and generating target networks based on the input data. Due to their versatility, HNs are especially useful for creating diverse personalized models.

To tackle the pFL challenge, a novel strategy called Personalised Federated Hypernetworks (pFedHN) [32] has been proposed. This approach utilizes hypernetworks, deep neural networks that generate the weights for another network based on their input. Hypernetworks are unique because they can learn multiple target networks simultaneously, allowing them to adapt to various specific needs. This makes them a promising tool in personalized FL and can be represented by equation 5. The updated pFL objective with HN is shown below:

$$\min_{\varphi, \boldsymbol{v}_1,\ldots,\boldsymbol{v}_n} F(\theta) = \frac{1}{|D_k|} \sum_{i \in D_k} f_i\left(h\left(\boldsymbol{v}_i; \varphi\right)\right) \tag{6}$$

where, $h\left(v_i; \varphi\right)$ is model-size independent, allowing any hyper-network size to enhance overall performance. Hyper-network updates are achieved through gradient chain rule calculations as below.

$$\nabla_\varphi f_i = \left(\nabla_\varphi \theta_i\right)^T \nabla_{\theta_i} f_i \tag{7}$$

To optimize θ_i and v_i, calculated $\nabla_{\theta_i} f_i$ and $\nabla_{\varphi} \theta_i$, respectively. The local client uses its training samples to optimize θ_i, determined by v_i and shared φ. This maintains uniqueness and enables parameter sharing without any contradictory local updates. This method trains a unified Heterogeneous Network (HN) model to generate a series of models, each tailored for an individual client. As you can see in Figure 2, this architecture allows for the sharing of parameters across clients while still being able to generate distinct and varied individual models.

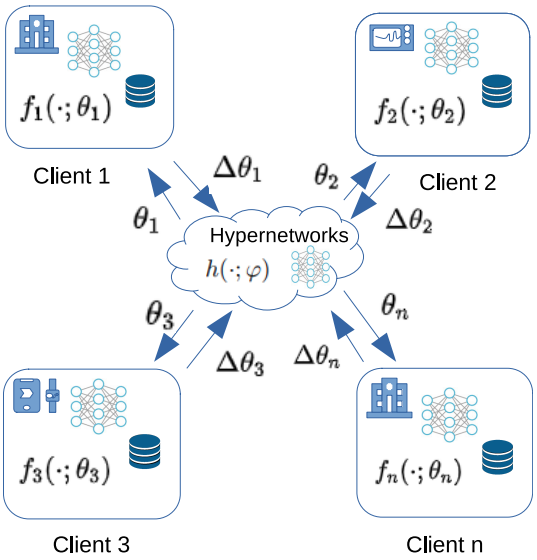

FIGURE 2. Personalized federated learning using hypernetwork framework in FedCure.

B. META-LEARNING BASED PFL

Meta-learning, or "learning to learn," improves learning by exposing the algorithm to various tasks. Optimization-based meta-learning algorithms are particularly useful as they can generalize and adapt quickly to new tasks. They are model-agnostic, making them applicable for supervised and reinforcement learning [33].

Federated Meta-Learning is an innovative approach combining FL and meta-learning principles to create highly personalized models for complex IoMT-based applications. In the context of IoMT networks, where diverse medical devices generate data with varying characteristics and follow unique data distributions, Federated Meta-Learning capitalizes on meta-learning to craft personalized models for individual devices. This process involves learning the unique characteristics of each device's data patterns and how they behave over time. Furthermore, Federated Meta-Learning facilitates knowledge transfer by training models on one device and adapting them to others with similar data patterns, thus building personalized models for devices with limited local data. IoMT environments can be optimized for edge computing to reduce the computational load and ensure scalability across various healthcare applications [26].

Per-FedAvg [15] an innovative method inspired by the popular FedAvg approach. Per-FedAvg is uniquely designed to address the pFL problem.

$$\min_{w \in \mathbb{R}^d} F(w) := \frac{1}{|D_k|} \sum_{i \in D_k} f_i \left(w - \alpha \nabla f_i(w)\right), \tag{8}$$

During each round of FedAvg, a certain number of users are chosen, and their models are updated through multiple gradient descent steps. On the other hand, Per-FedAvg also involves selecting users. However, the focus is on personalizing the solution for the equation mentioned in 8. It's worth noting that this equation can be considered an average of individual "meta-functions," each linked to a specific user. These meta-functions are defined to adapt models to their respective data and loss functions. This personalized approach improves the effectiveness of FL, especially in environments like the IoMT. Compared to the Federated Transfer Learning strategy, the Federated Meta Learning approach is more difficult to deploy since it frequently uses complex training algorithms. On the other hand, Federated Meta Learning produces a more reliable model, which can be especially helpful for devices with small data sets.

C. REGULARIZATION BASED PFL

ML models are susceptible to overfitting, resulting in subpar performance when confronted with new, unseen data. Regularization techniques are frequently employed to combat this challenge during the model training process. In FL, a distributed learning paradigm where data is distributed across diverse devices, regularization can be harnessed to control the influence of local updates on the global model. This, in turn, enhances the stability of convergence and the overall generalization of the global model, ultimately yielding improved personalized models.

Each client device in FL has its own local data and model parameters. To update the global model, these local models need to be combined to minimize a global objective function. Typically, each client minimizes its local objective function, a function of its local model parameters and data. However, this can lead to overfitting, especially when the local data is limited or noisy. Regularization can be applied to the local objective functions to prevent overfitting and improve generalization. Specifically, each client minimizes the following regularized objective function:

$$\min_{\theta \in \mathbb{R}^d} h_k(\theta; w) := f_k(\theta) + l_{\text{reg}}(\theta; w) \tag{9}$$

Here, $f_k(\theta)$ is the local objective function of client k. The regularization term $l_{\text{reg}}(\theta; w)$ penalizes complex models and encourages simpler models that generalize better. This function is contingent on the model parameters θ and the global model parameters w. Regularisation of the local objective functions makes it possible to limit the impact of local updates on the global model, improving convergence stability and strengthening the global model's ability to generalize. For a visual representation, please refer to

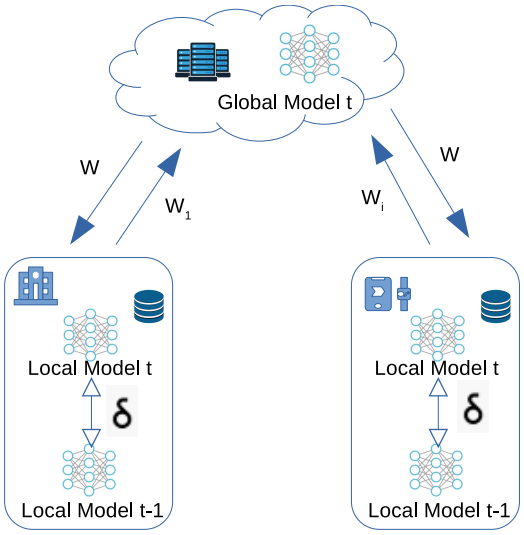

FIGURE 3. Personalized federated learning via regularization in FedCure.

Figure 3, which provides an overview of achieving model personalization through regularizing local losses.

Recently, a novel approach known as MOON (Minimizing Weight Divergence and Optimizing for Fast Convergence in FL) has emerged within the field of FL [34]. The main goal of MOON is to reduce the difference between the global model and the learned representations of local models. It does this by focusing on weight divergence, a metric that measures the difference in weights between local and global models. Beyond minimizing weight divergence, MOON also prioritizes expediting the convergence process in FL. To this end, the difference between the representations learned by a particular local model and its prior iteration is amplified. This incentivizes the local model to learn enhanced representations and progress from its prior version, thus accelerating the learning trajectory. MOON boasts multiple advantages, starting with each client's ability to acquire a representation that closely aligns with the global model, thereby reducing local model discrepancies. Additionally, it stimulates local models to learn more refined representations compared to their earlier versions, resulting in a faster learning process. Overall, the MOON approach can potentially enhance the efficiency and effectiveness of FL, rendering it a more proficient technique for a wide array of machine-learning tasks. Regularization can be applied in various ways in FL, depending on the type of regularization used and the optimization algorithm used to solve the global objective function [33]. L1 and L2 regularisation, dropout, and early halting are a few regularisation strategies applied in FL.

Regularisation inside FL can improve the performance of the global model and, in the end, lead to the development of more customized models for each client device. Utilizing the similarity between model representations is the core idea behind MOON regarding IoMT edge devices, hospitals, and medical facilities working together with a cloud-based FL server. This is done to improve each entity's local training procedures.

D. MULTI-TASK LEARNING BASED PFL

Multi-task learning is a powerful technique in IoMT-based healthcare applications to address statistical heterogeneity and foster relationship modeling among devices. This approach allows models to learn from different data sources simultaneously, making them highly adaptable to the heterogeneity of IoMT data. Additionally, it aids in identifying underlying relationships between clients, making it a powerful tool for personalization. With the help of multi-task learning, we can learn several tasks simultaneously, utilizing the shared knowledge to enhance our performance on each task. FL aims to create a shared model across IoMT devices. Conversely, federated multi-task learning focuses on distinct tasks across various devices and seeks to unveil inherent model relationships while safeguarding privacy. This approach allows each device's model to gain insights from the information gleaned from other devices, resulting in a device-specific model that is consistently personalized to suit the unique characteristics of each device.

Figure 4 shows that Federated multi-task learning uses model parameters from the edge and IoMT devices to give the cloud server insights into the relationships between different tasks. Afterward, any device can modify its parameters based on its information and the current relationships between the models. This collaborative approach empowers healthcare devices to collectively train local models, effectively addressing statistical differences and creating high-quality personalized models.

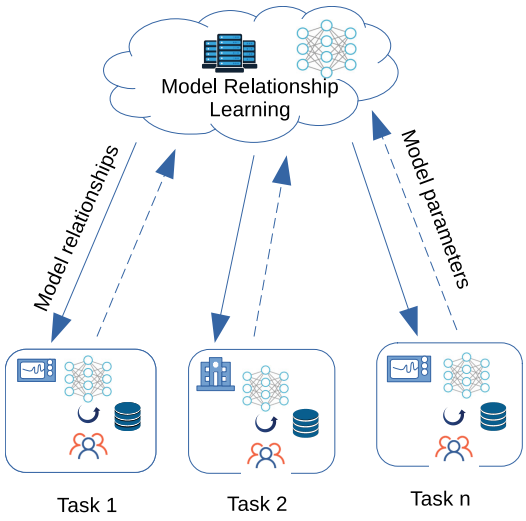

FIGURE 4. Multi-task federated learning for FedCure.

FedAMP [35] represents an innovative FL framework designed to foster cooperation among clients who share similar data distributions. This architecture creates and maintains a customized cloud model on the central server for each client, which is put together by linearly combining

local models. The customer receives the customized cloud model for localized training at the end of each communication cycle. The local model's weights are obtained by optimizing a certain objective function, as explained in the source.

$$\theta_k^* = \arg\min_{\theta \in \mathbb{R}^d} f_k(\theta) + \frac{\mu}{2\alpha}||\theta - u_k||^2 \tag{10}$$

Here, α represents the step size of the gradient descent. FedAMP's unique approach ensures more personalized and effective model training, promoting stronger client collaboration. It can improve FL systems' performance, particularly among clients with similar data distributions. This paper adopted MOCHA [36], a distributed optimization method used in complex IoMT network-based healthcare applications. MOCHA optimizes communication efficiency, reduces communication rounds, and minimizes the impact of stragglers. Furthermore, it introduces an asynchronous updating approach and demonstrates resilience in fault tolerance issues. Nevertheless, the conventional federated multi-task learning approach encounters certain constraints in IoMT scenarios due to the inherent device disparities. Exploring cluster-based federated multi-task learning could offer a promising avenue for future research in this domain.

E. MODEL INTERPOLATION BASED PFL

Model Interpolation is an advanced technique used in personalized FL [37]. This technique balances personalization and generalization, as shown in Figure 5. This is achieved by blending a client's local model, represented by M_{edge}, with a global model shared across all clients, represented by M_{server} and P_m denoted as a personalized model. A parameter called λ is used to adjust the degree of personalization. When λ is set close to 1, the combined model leans heavily towards the local model, emphasizing the uniqueness of the client's data. Conversely, when λ approaches 0, the influence of the global model dominates, emphasizing the common patterns shared across all clients.

$$P_m = \lambda \cdot M_{edge} + (1 - \lambda) \cdot M_{server} \tag{11}$$

FL is an efficient approach that can be integrated into healthcare applications. It maintains communication cost and security levels while training a single model. In IoMT environments, patient data is distributed across various healthcare providers and medical devices. FL can be used along with model interpolation to create personalized healthcare models for individual patients to ensure privacy and data security. FL faces the challenge of combining global and local models while adapting to unique communication constraints. The APFL algorithm introduces a dynamically learned mixing parameter for each client. In contrast, the HeteroFL framework trains local models with diverse complexities while operating based on a single global model. A practical and resource-efficient approach for implementing model interpolation in FL utilizes the Bagging algorithm to combine multiple models and enhance generalization [38].

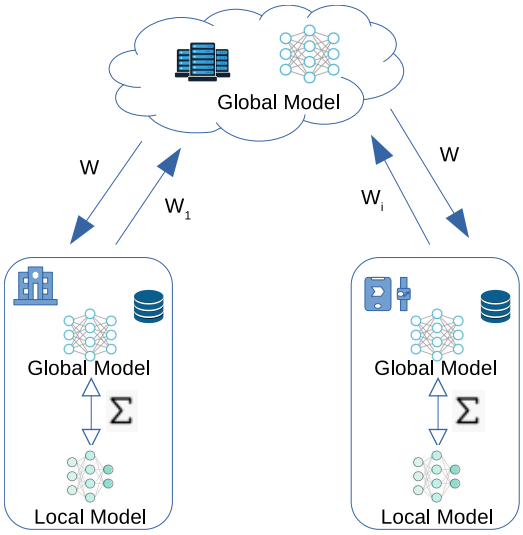

FIGURE 5. Model interpolation-based Personalized federated learning for FedCure.

They use shallow neural networks as basic learners to minimize computing resources and communication bandwidth requirements. The approach enables information exchange among various client clusters and achieves personalized FL. The proposed approach is promising for enhancing FL in scenarios with resource constraints and heterogeneous data distributions. These tools empower the development of tailored and effective healthcare solutions that respect each client's distinct characteristics and constraints. FL and model interpolation are powerful techniques that can create personalized healthcare models while preserving data privacy and security.

F. KNOWLEDGE DISTILLATION BASED PFL

In IoMT environments, entities like hospitals, research institutions, and IoMT devices face unique challenges when engaged in FL. Unlike traditional FL, healthcare and medical research participants possess the capacity and inclination to develop their own distinct ML model due to data privacy and intellectual property concerns. The heterogeneity of models in IoMT environments presents a challenge to conventional FL practices, requiring a focus on accommodating diverse model architectures while deriving collective insights from their data. The variety of models involved in the process can create difficulties for regular FL, posing new challenges that must be addressed.

Knowledge Distillation (KD) [39] is a popular method for transferring valuable insights from a set of teacher models to a more lightweight student model in FL. KD typically involves representing knowledge as class scores. There are four principal architectures for FL-based knowledge distillation. The client-centric approach focuses on distilling knowledge independently to each FL client, enabling them to refine their personalized models by assimilating collective wisdom from the teacher models. The server-centric approach

directs the knowledge distillation process toward bolstering the central FL server's model by incorporating insights from the teacher models. Bidirectional knowledge distillation facilitates knowledge exchange between clients and the server to enhance the FL ecosystem comprehensively. Lastly, inter-client knowledge distillation promotes collaborative learning among FL clients, where they mutually share and transfer knowledge, fostering cooperative model enhancement. These distinct KD strategies are pivotal in advancing the effectiveness and resilience of FL models tailored to diverse applications and use cases.

To tackle the issue of model heterogeneity in FL, clients can train multiple models utilizing their private data through KD [40]. This procedure calculates a consensus using a public dataset's average class ratings. Every client uses its private dataset to fine-tune its model after each communication round, using the public datasets with the latest consensus. Using the combined expertise of other clients, this method allows each client to receive a customized model. Additionally, FedGen [41], presents a framework for data-free distillation that uses a generative model trained on the FL server to provide FL clients with knowledge. Clients utilize this knowledge as an inductive bias to generate augmented representations and regulate their local learning process. Additionally, the FedDF algorithm [42] acknowledges that edge clients may require different model architectures due to their varying processing capacities. Here, the FL server creates multiple distinct prototype models. Cross-architecture learning is facilitated through ensemble distillation, employing an unlabeled public dataset to train each student model. KD proves to be a versatile technique for bidirectional exchanges, enhancing FL in diverse contexts. He et al. [8] proposed FedGKT, which uses bidirectional distillation and alternating lowering to train tiny edge models and a bigger server model. This optimizes computing by moving the computational load from edge clients to the more potent FL server. By controlling its local loss, this technique is a distributed algorithm for on-device learning that continually updates the client's model weights. By sharing information with nearby FL clients in the network, this cyclical process facilitates distributed and collaborative learning and speeds up model learning. Participants in IoMT-based healthcare applications can effectively address issues resulting from heterogeneous model architectures by implementing Knowledge Distillation. With this approach, participants can use high-performance, personalized machine-learning models tailored to their particular operational and data requirements. As a result, it improves healthcare decision-making, facilitates medical research, and improves patient care.

V. CASE STUDY

This section discusses various use cases tested, the datasets utilized, and the pFL approaches supported, all within different heterogeneity settings. This analysis is conducted using the proposed framework, FedCure. The experiments considered the use case of diabetes monitoring, remote health monitoring, maternal healthcare, eye retinopathy classification, and Human Activity Recognition (HAR) with publicly available datasets.

Table 1 provides a comprehensive summary of the FedCure framework's compatibility with various pFL approaches and algorithms in the context of distinct healthcare datasets. FL Approaches lists different FL strategies specifically designed for various healthcare use cases. pFL Algorithms introduce pFL algorithms customized to meet the unique requirements of different healthcare applications. These algorithms allow for personalized model training within the FedCure framework. The rest of the table represents various healthcare datasets, such as Diabetes,[1] Body Performance,[2] Maternal Health,[3] OCT Images,[4] UCI HAR,[5] and PAMAP2.[6] For each dataset, the table indicates whether the associated FL approach or pFL algorithm is supported by displaying "Yes" or "No."

A. DATASETS AND IMPLEMENTATION DETAILS

The datasets used in this study form the backbone of various healthcare applications, each tailored to address specific healthcare challenges. They are pivotal for conducting experiments using FedCure with different healthcare use cases. Table 2 shows the datasets and their relevance within the context of specific healthcare applications. <Diabetes Dataset> Diabetes Monitoring: This dataset, curated for diabetes monitoring, encompasses 769 samples with eight attributes. It focuses on a binary classification task related to diabetes. In the FL experiments, 20 clients contributed their data, reflecting real-world healthcare scenarios. <Body Performance Dataset> Remote Health Monitoring (RHM): The body performance dataset is employed for RHM, a sample size of 13,394 samples with eleven attributes. It covers a classification problem with four distinct classes. Similar to the diabetes dataset, 20 clients are involved, mirroring remote health monitoring situations. <Maternal Health Datase> (Maternal Health Care): Specifically designed for maternal health care, this dataset consists of 1,015 samples and six attributes. It tackles a three-class classification problem and, once again, involves 20 clients in the FL process. <OCT Image Dataset> (Eye Retinopathy Classification): Eye retinopathy classification hinges on a remarkable scale dataset comprising 84,495 samples, with images at 256 × 256 pixel resolution. The task involves classifying retinopathy into four categories. Here, 10 clients partake in FL experiments. <UCI HAR Dataset> (Human Activity Recognition): With a dataset of 10,299 samples, the HAR dataset includes nine attributes for recognizing human activities across six categories. This dataset leverages

[1] https://www.kaggle.com/datasets/mathchi/diabetes-data-set
[2] https://www.kaggle.com/datasets/kukuroo3/body-performance-data
[3] https://www.kaggle.com/datasets/csafrit2/maternal-health-risk-data
[4] https://www.kaggle.com/datasets/paultimothymooney/kermany2018/
[5] https://archive.ics.uci.edu/dataset/240/human+activity+recognition
[6] https://archive.ics.uci.edu/dataset/231/pamap2+physical+activity+monitoring

TABLE 1. FedCure supported datasets and pFL approaches.

FL Approaches	pFL Algorithms	Datasets Supported for Usecases					
		Diabetes	Maternal Health	Body Performance	OCT Images	UCI HAR	PAMAP2
Traditional FL	FedAvg [9]	Yes	Yes	Yes	No	Yes	Yes
Model-Regularization-based pFL	MOON [34]	Yes	Yes	Yes	No	Yes	Yes
Multi-Task based pFL	FedMTL [36]	Yes	Yes	No	No	Yes	Yes
Hyper Network based pFL	pFedHN [32]	No	No	No	Yes	No	No
Knowledge Distillation based pFL	FedDistill [40], FedProto [41], FedPAC [42].	Yes	Yes	Yes	No	Yes	Yes
Model-Interpolation based pFL	FedPer [38]	No	No	Yes	No	No	No
Meta-Learning based pFL	Per-FedAvg [15]	Yes	No	No	No	No	No

TABLE 2. Details about datasets used.

Dataset	Use cases	Sample Size	Attributes	Classes	Clients
Diabetes	Diabetes Monitoring	769	8	2	20
Body Performance	Remote Health Monitoring	13394	11	4	20
Maternal Health	Maternal Health Care	1015	6	3	20
OCT image	Eye retinopathy classification	84495	256*256	4	10
UCI HAR	Human Activity Recognition	10299	9*1*128	6	18
PAMAP2	Human Activity Recognition	15012	9*3*256	12	9

TABLE 3. FedCure supported heterogeneity setting.

Data Heterogeneity Setting		System Heterogeneity Setting		
Pathological Non -IID and Unbalanced	Practical Non-IID and unbalanced	Struggler effect	Slow learner	Slow Sender

18 clients in the FL process. <PAMAP2 Dataset> (Human Activity Recognition): Another dataset for HAR, PAMAP2, contains 15,012 samples. It presents nine attributes, classifying activities into 12 unique categories. In this FL setup, nine clients actively participate. In order to assess the effectiveness of the FedCure framework, it is crucial to carefully divide all datasets into separate training and testing sets. The recommended approach involves assigning 80% of the data for training purposes while setting aside the remaining 20% for evaluating the model's ability to generalize. These datasets are crucial for training and evaluating Federated Learning (FL) models, with each dataset serving a unique purpose in healthcare applications. Collectively, they reflect the diversity and intricacy of real-world healthcare data, making them indispensable resources for developing and evaluating FL solutions within the healthcare domain.

Table 3 demonstrates how FedCure is adaptable and robust in addressing real-world challenges by presenting the various heterogeneity settings explored in the framework. FedCure is designed to handle diverse data and system heterogeneity scenarios, ensuring its effectiveness across practical applications. To address data heterogeneity issues, FedCure considers two distinct situations: pathological non-IID and practical non-IID scenarios, each reflecting unique challenges. In the pathological non-IID scenario, the individual clients' data is characterized by an extreme form of non-IID distribution, such as clients possessing data with only a specific subset of labels, even though the complete dataset encompasses a broader spectrum of categories. In contrast, in the practical non-IID scenario, FedCure uses the Dirichlet distribution [12] to simulate realistic non-IID data in healthcare applications, effectively modeling variations in data distribution across different clients. FedCure addresses different key factors, including the struggler effect, slow learners, and slow senders, all contributing to the dynamic nature of FL in healthcare applications. FedCure employs a dropout rate to address the struggler effect, where selected clients are randomly dropped at each training round. Clients designated as "slow trainers" persistently train at a slower pace than their peers, while clients identified as "slow senders" consistently make their data slower. These measures ensure FedCure adapts and manages the varying system heterogeneity within healthcare FL. FedCure's adaptability across these diverse settings underscores its efficacy in addressing the complexity of healthcare-related FL.

The comprehensive overview of the models utilized for different case studies is detailed below. For the diabetes dataset, a Logistic Regression model is employed, taking 8 attributes as input and generating a binary output using a softmax activation function. In the case of the maternal health dataset, a Deep Neural Network (DNN) architecture is adopted, consisting of one input layer with 6 neurons, one hidden layer with 20 neurons, and an output layer with 3 neurons. The body performance dataset also utilizes a DNN with a similar architecture, featuring one input layer with 11 neurons, one hidden layer with 20 neurons, and an output layer with 4 neurons. These DNN models employ softmax activation at the output layer and categorical cross-entropy as the loss function. On the other hand, a more complex Custom Convolutional Neural Network (CNN) is designed for the OCT Images dataset. It incorporates two convolutional layers with distinct hyperparameters and max-pooling layers

with specific stride lengths. Three fully connected layers are also included, where the output layer consists of 4 classes for classification. The first convolution layer consists of 3 input channels, 16 output channels, kernel size 5, and the second convolution layer is defined as a 16 input channels, 32 output channels, kernel size 5. These tailored models are crafted to address the unique characteristics of each dataset, ensuring optimal performance for their respective tasks. For the HAR dataset, a CNN architecture has been implemented. This architecture consists of two convolutional layers. The first convolution layer uses hyperparameters with input channels set to 9, output channels to 32, and a kernel size of (1,9), and applies the Rectified Linear Unit (ReLU) activation function. A max-pooling layer with a kernel size of (1,2) and a stride of 2 is employed. The second convolution layer has hyperparameters: input channels 32, output channels 64, kernel size is (1,9), and ReLU activation. Again, a max-pooling layer follows with the same kernel size and stride. The fully connected layers in this network consist of an input layer with 1664 neurons, two hidden layers with 1024 and 512 neurons, and an output layer with six neurons to classify the six different activities. Similarly, for the PAMAP2 dataset, a CNN architecture is adopted. This architecture mirrors the HAR model with two convolutional layers, max-pooling layers, and fully connected layers. However, the PAMAP2 dataset's CNN differs in terms of the input layer configuration, which consists of 3712 neurons, and it has an output layer with 12 neurons, corresponding to the 12 different classes for classification. These CNN architectures are specifically designed to capture and classify patterns in sensor data from the respective datasets for activities and action recognition.

B. EXPERIMENTS RESULTS AND ANALYSIS

This work conducted experiments on five different case studies, namely Diabetes monitoring, Human Activity Recognition, Eye Retinopathy Classification, Remote Healthcare, and Maternal health monitoring. These case studies utilized different sizes of datasets with two types of heterogeneity settings: pathological non-IID unbalanced and practical non-IID unbalanced. The experiments were simulated multiple times to ensure a fair study. Accuracy is a fundamental metric that measures the overall correctness of the model predictions. In the context of intelligent IoMT applications, accuracy provides a clear indication of how well FedCure is able to correctly classify instances. This metric is essential for evaluating the general efficacy of the model in making accurate predictions across different tasks. The results were compared with centralized, traditional FL, and pFL approaches. The centralized approach involved using a single system with complete data. On the other hand, in the FL approach, each application case study divided the data set into different clients, as mentioned in Table 2. Five different approaches were adopted for personalized FL, as shown in Table 1. In the pFL approach, individual clients could tailor their models to suit their needs. Notably, clients also had the option to delegate their learning tasks from their devices to nearby edge computing resources, like an edge server located in the hospital. The proposed FedCure framework allows for swift and efficient computation. Furthermore, practical device heterogeneity like client dropout, slow sender, and slow learner was also included in the experiments, as shown in Tables 2 and 3. This helped to create a more realistic and practical scenario, as these types of heterogeneities are common in real-world applications.

In the context of the Diabetes Monitoring case study and Remote Health monitoring, we evaluated the experiment's performance under two different data heterogeneity settings involving 20 clients. In real-world IoMT-based FL healthcare applications, it's common for clients not to participate in every communication round. To mimic this scenario, we conducted experiments to assess the effectiveness of our proposed FedCure framework. This framework supports a client dropout ratio of 20%, and in another scenario, 10% of clients are slower learners and slower data senders. Our experiments spanned 2,500 communication rounds and contained the traditional FL (FedAvg) approach and various personalized pFL approaches supported by FedCure. These pFL approaches included KD, Multi-Task, Regularization, and Meta-Learning-based methods. Remarkably, the results highlighted the superior performance of pFL approaches over traditional FL (FedAvg) in the face of both heterogeneous environmental settings and observed in Diabetes Monitoring, pFL approach KD (FedProto) even below the performance of traditional FL. It's essential to emphasize that different application case studies demand tailored approaches, as there isn't a one-size-fits-all solution in pFL. Among these approaches, the multi-task-based pFL approach demonstrated the highest accuracy in pathological and practical non-IID unbalanced settings of the Diabetes Monitoring case study, as illustrated in Figure 6. In Remote Health Monitoring, all pFL approaches perform well, as shown in Figure 7. This outcome holds significant implications for guiding the development of IoMT-based FL healthcare applications, showing that a thoughtful choice of pFL strategy can yield substantial benefits in real-world healthcare scenarios.

In the Maternal Health monitoring case study, we maintained consistency with the experimental approach employed in the Diabetes and Remote Health Monitoring studies. With a group of 20 clients, we executed 2,500 communication rounds, utilizing pathological and practical unbalanced non-IID data for our investigations. This uniformity in our experimental setup allowed us to draw meaningful comparisons and comprehensively evaluate our methodologies. Figure 8 illustrates how various approaches perform regarding accuracy. These experiments reaffirmed that pFL approaches outperform traditional FL in handling the diverse environments encountered in IoMT data-based applications. In the case of pathological non-IID data distributions, we noticed significant fluctuations in the learning process. This signifies that data on different clients exhibits variations in quality, quantity, and distribution, impacting the learning

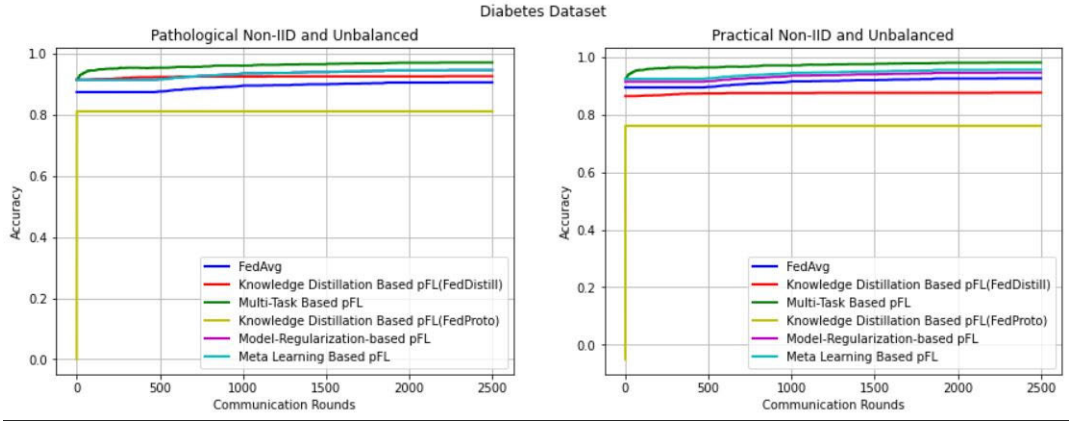

FIGURE 6. Accuracy vs. communications rounds of diabetes monitoring case study.

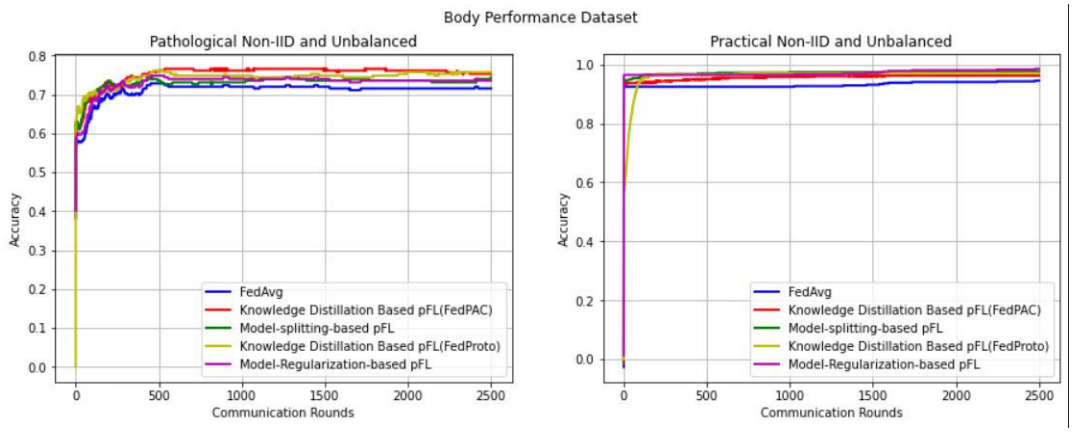

FIGURE 7. Accuracy vs. communications rounds of remote health monitoring case study.

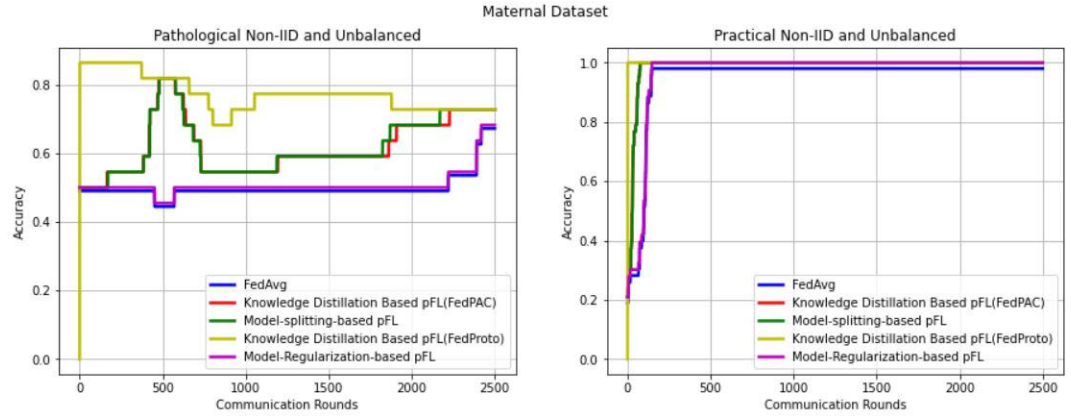

FIGURE 8. Accuracy vs. communications rounds of maternal health monitoring case study.

process at each round of FL in practical IoMT-based environments. As the number of communication rounds increases, the model's learning improves. Still, it is worth noting that in the KD-based pFL algorithm, FedProto's performance tends to decline, mirroring our findings in the Diabetes monitoring case study. On the other hand, in practical non-IID settings, all pFL approaches prove to be more effective than traditional FL approaches.

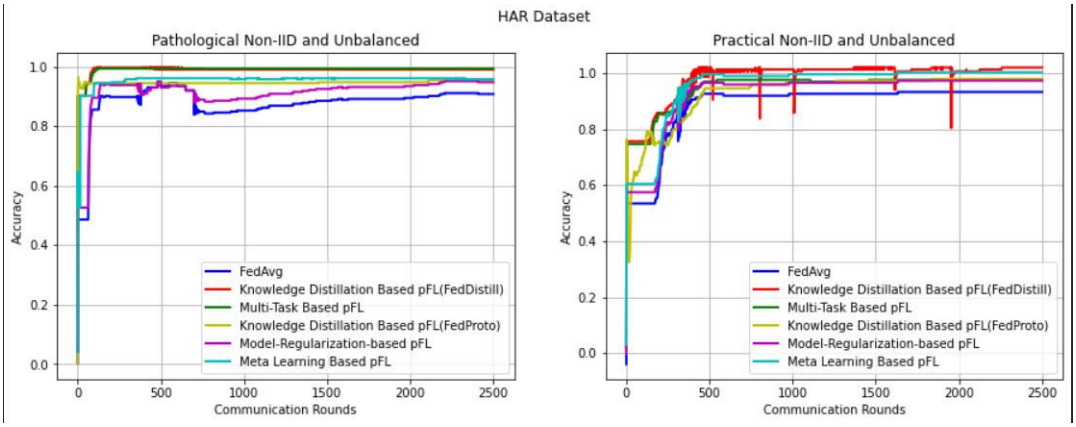

FIGURE 9. Accuracy vs. communications rounds of har case study.

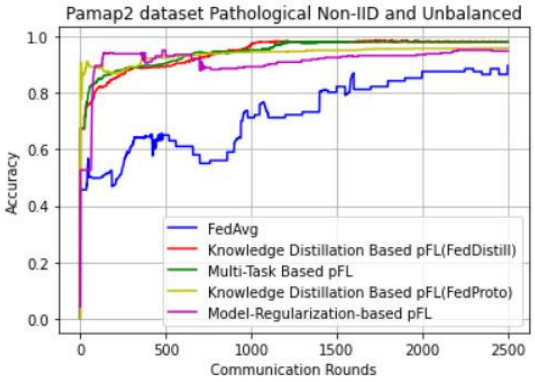

FIGURE 10. Accuracy vs. Communications rounds of HAR (PAMP2) case study.

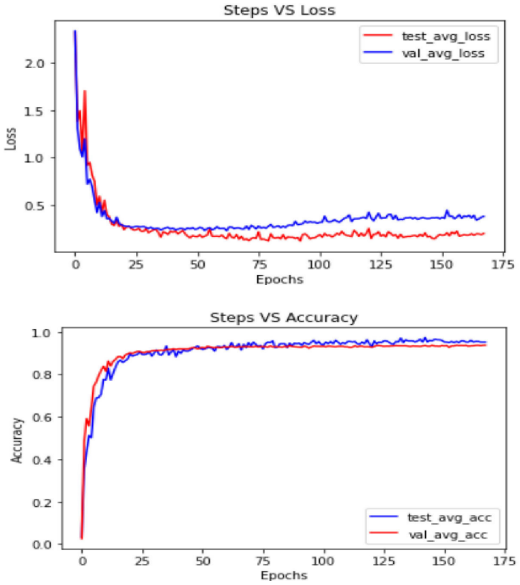

FIGURE 11. Performance analysis of eye retinopathy classification using hypernetwroks.

In the HAR case study, we employed two widely accessible datasets. We maintained the same device heterogeneity settings as those used in the aforementioned case studies. We examined the UCI HAR dataset in both pathological and practical non-IID data contexts, involving 18 clients. For the PAMAP2 dataset, we tested only the practical non-IID data setting with 9 clients due to data quantity considerations. This case study provides further compelling evidence of the superior performance of pFL approaches compared to traditional FL methods. Figure 9 illustrates the accuracy of the UCI HAR dataset, where KD-based pFL notably outperforms other approaches. It's important to note that while KD-based pFL excels in this setting, it exhibited decreased performance in other case studies, highlighting the need for different approaches in varying heterogeneous environments. This study underscores the importance of adapting to different pFL approaches to train optimal models. The proposed FedCure framework offers flexibility texpanding these experiments to encompass a broader range of scenarios and applications would be valuable training. Figure 10 showcases the performance accuracy of the PAMAP2 dataset in a pathological non-IID unbalanced data setting. Collectively, these case studies emphasize the versatility and effectiveness of pFL in addressing heterogeneity issues within FL across diverse healthcare applications. In the context of Eye Retinopathy Classification, we worked with an image dataset that underlines the effectiveness of pFL approaches in cross-silo FL scenarios. This setting reflects a scenario where multiple medical organizations collaborate to develop a shared model for performing Eye Retinopathy Classification. We opted for a non-heterogeneous setting for this particular case study, specifically a non-IID balanced dataset involving 10 clients. Notably, we didn't introduce device heterogeneity into this experiment since the participating organizations boasted substantial computational power and robust communication capabilities. The results, depicted in Figure 11, highlight

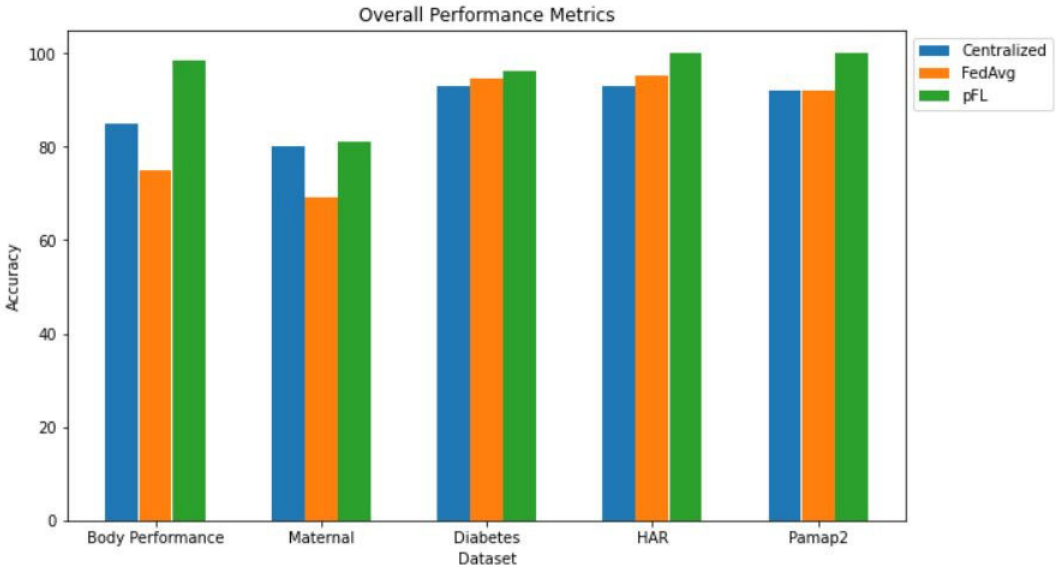

FIGURE 12. Accuracy of different learning approaches of all case studies tested.

the exceptional performance of hypernetwork-based pFL in handling image-based datasets. This underscores the potential of hypernetwork-based approaches for scenarios like eye retinopathy classification, where image data plays a crucial role. In future research endeavors, it would be valuable to expand these experiments to encompass a broader range of scenarios and applications, exploring the full potential of hypernetwork-based pFL across different healthcare contexts.

Figure 12 provides a comprehensive performance overview across all case studies, excluding Eye Retinopathy Classification. It showcases the performance accuracy achieved through centralized training, traditional FL, and the best-performing pFL approach for each case study in a pathological non-IID data setting. Comparing these results shows that FL can deliver strong performance while preserving data privacy. With personalization, where each client fine-tunes the model with its unique data, we witness minimal accuracy variations among clients. This observation underscores the power of personalization, as it captures fine-grained personal information. The personalized models for each participant help mitigate the performance degradation associated with non-IID data distributions. These experiments consistently highlight the superiority of pFL in heterogeneous IoMT network environments. They underscore the enhanced performance and data privacy advantages personalized FL offers compared to traditional approaches. This finding holds significant promise for complex IoMT networks and healthcare applications.

The experiments provide compelling evidence for the effectiveness of FedCure, a heterogeneity-aware Personalized FL framework. This framework demonstrates its ability to construct efficient models for intelligent IoMT-based healthcare applications within complex networks. The framework encompasses five distinct case studies and six healthcare datasets, incorporating five pFL approaches. It excels in handling two data heterogeneity settings, making it a versatile and scalable solution for real-time applications in the healthcare domain. The FedCure framework has its strengths, but there are also some limitations that need to be addressed. The use of edge computing in FedCure raises concerns about the security and privacy of sensitive health data, particularly with regards to privacy concerns. Edge devices may be vulnerable to local attacks, and the transmission of personalized model updates between edge and cloud components could expose potential vulnerabilities. It is important to ensure that robust security measures and encryption protocols are in place to safeguard patient data. Additionally, the framework should address potential privacy risks associated with sharing global models across devices, even in an FL setting, to mitigate the risk of unintended data exposure.

VI. CONCLUSION

This paper presents FedCure, an innovative personalized FL framework tailored for intelligent IoMT applications, delivering robust data privacy protection in a cloud-edge architecture. FedCure empowers the acquisition of a globally shared model by amalgamating local updates sourced from distributed IoMT devices, effectively capitalizing on edge computing capabilities. Notably, it tackles the inherent heterogeneities encompassing device disparities, statistical variations, and model diversity in IoMT environments. This comprehensive approach integrates diverse personalized FL techniques to achieve tailored personalization and elevate the performance of individual devices. The case studies,

spanning human activity recognition, Eye Retinopathy classification, diabetes, Maternal, and Remote Health Monitoring, underscore FedCure's potential to cater to a wide range of intelligent IoMT applications. Future research endeavors will expand the horizons by conducting additional experiments, especially within the domain of Hypernetwork-based pFL for tasks such as medical image segmentation and classification. Moreover, the ongoing development of the FedCure framework will include integrating additional pFL approaches and diverse datasets to enhance its capabilities further.

REFERENCES

[1] S. A. Khowaja, A. G. Prabono, F. Setiawan, B. N. Yahya, and S.-L. Lee, "Contextual activity based healthcare Internet of Things, services, and people (HIoTSP): An architectural framework for healthcare monitoring using wearable sensors," *Comput. Netw.*, vol. 145, pp. 190–206, Nov. 2018.

[2] V. Shah and A. Khang, "Internet of Medical Things (IoMT) driving the digital transformation of the healthcare sector," in *Data-Centric AI Solutions and Emerging Technologies in the Healthcare Ecosystem*. Boca Raton, FL, USA: CRC Press, 2023, pp. 15–26.

[3] Z. Ashfaq, A. Rafay, R. Mumtaz, S. M. H. Zaidi, H. Saleem, S. A. R. Zaidi, S. Mumtaz, and A. Haque, "A review of enabling technologies for Internet of Medical Things (IoMT) ecosystem," *Ain Shams Eng. J.*, vol. 13, no. 4, Jun. 2022, Art. no. 101660.

[4] C.-H. Lin, H.-Y. Lai, P.-T. Huang, P.-Y. Chen, N.-S. Pai, and F.-Z. Zhang, "Combining riemann-lebesgue based key generator and machine learning based intelligent encryption scheme for IoMT images infosecurity," *IEEE Internet Things J.*, vol. 11, no. 1, pp. 1344–1360, Jan. 2024.

[5] M. F. Khan, T. M. Ghazal, R. A. Said, A. Fatima, S. Abbas, M. A. Khan, G. F. Issa, M. Ahmad, and M. A. Khan, "An IoMT-enabled smart healthcare model to monitor elderly people using machine learning technique," *Comput. Intell. Neurosci.*, vol. 2021, pp. 1–10, Nov. 2021.

[6] P. Voigt and A. Von dem Bussche, "The EU general data protection regulation (GDPR)," in *A Practical Guide*, vol. 10, 1st ed. Cham, Switzerland: Springer, 2017, p. 5555.

[7] S. D. Calloway and L. M. Venegas, "The new HIPAA law on privacy and confidentiality," *Nursing Admin. Quart.*, vol. 26, no. 4, pp. 40–54, 2002.

[8] C. He, M. Annavaram, and S. Avestimehr, "Group knowledge transfer: Federated learning of large CNNs at the edge," in *Proc. NIPS*, vol. 33, 2020, pp. 14068–14080.

[9] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. Y. Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Proc. 20th Int. Conf. Artif. Intell. Statist.*, 2017, pp. 1273–1282.

[10] P. Kairouz, H. B. McMahan, B. Avent, A. Bellet, M. Bennis, A. N. Bhagoji, K. Bonawitz, Z. Charles, G. Cormode, and R. Cummings, "Advances and open problems in federated learning," *Found. Trends Mach. Learn.*, vol. 14, nos. 1–2, pp. 1–210, 2021.

[11] J. Wang, Q. Liu, H. Liang, G. Joshi, and H. V. Poor, "Tackling the objective inconsistency problem in heterogeneous federated optimization," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 7611–7623.

[12] T. Li, A. K. Sahu, M. Zaheer, M. Sanjabi, A. Talwalkar, and V. Smith, "Federated optimization in heterogeneous networks," in *Proc. Mach. Learn. Syst.*, vol. 2, 2020, pp. 429–450.

[13] Y. Zhao, M. Li, L. Lai, N. Suda, D. Civin, and V. Chandra, "Federated learning with non-IID data," 2018, *arXiv:1806.00582*.

[14] A. Rakotomamonjy, M. Vono, H. Jesse Medina Ruiz, and L. Ralaivola, "Personalised federated learning on heterogeneous feature spaces," 2023, *arXiv:2301.11447*.

[15] A. Fallah, A. Mokhtari, and A. Ozdaglar, "Personalized federated learning with theoretical guarantees: A model-agnostic meta-learning approach," in *Proc. NIPS Conf.*, vol. 33, Dec. 2020, pp. 3557–3568.

[16] C. T. Dinh, N. Tran, and J. Nguyen, "Personalized federated learning with Moreau envelopes," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 21394–21405.

[17] A. Gupta, S. Misra, N. Pathak, and D. Das, "FedCare: Federated learning for resource-constrained healthcare devices in IoMT system," *IEEE Trans. Computat. Social Syst.*, vol. 10, no. 4, pp. 1587–1596, Aug. 2023.

[18] S. F. Ahmed, M. S. B. Alam, S. Afrin, S. J. Rafa, N. Rafa, and A. H. Gandomi, "Insights into Internet of Medical Things (IoMT): Data fusion, security issues and potential solutions," *Inf. Fusion*, vol. 102, Feb. 2024, Art. no. 102060.

[19] S. C. Sethuraman, P. Kompally, S. P. Mohanty, and U. Choppali, "MyWear: A novel smart garment for automatic continuous vital monitoring," *IEEE Trans. Consum. Electron.*, vol. 67, no. 3, pp. 214–222, Aug. 2021.

[20] L. Rachakonda, S. P. Mohanty, and E. Kougianos, "ILog: An intelligent device for automatic food intake monitoring and stress detection in the IoMT," *IEEE Trans. Consum. Electron.*, vol. 66, no. 2, pp. 115–124, May 2020.

[21] Q. Wu, X. Chen, Z. Zhou, and J. Zhang, "FedHome: Cloud-edge based personalized federated learning for in-home health monitoring," *IEEE Trans. Mobile Comput.*, vol. 21, no. 8, pp. 2818–2832, Aug. 2022.

[22] J. A. Alzubi, O. A. Alzubi, A. Singh, and M. Ramachandran, "Cloud-IIoT-based electronic health record privacy-preserving by CNN and blockchain-enabled federated learning," *IEEE Trans. Ind. Informat.*, vol. 19, no. 1, pp. 1080–1087, Jan. 2023.

[23] W. Lu, J. Wang, Y. Chen, X. Qin, R. Xu, D. Dimitriadis, and T. Qin, "Personalized federated learning with adaptive batchnorm for healthcare," *IEEE Trans. Big Data*, early access, May 23, 2022, doi: 10.1109/TBDATA.2022.3177197.

[24] Q. Xia, W. Ye, Z. Tao, J. Wu, and Q. Li, "A survey of federated learning for edge computing: Research problems and solutions," *High-Confidence Comput.*, vol. 1, no. 1, Jun. 2021, Art. no. 100008.

[25] S. Alam, "Federated learning benchmarks and frameworks for artificial intelligence of things," Ph.D. dissertation, Dept. Comput. Sci., Michigan State Univ., East Lansing, MI, USA, 2023.

[26] Q. Wu, K. He, and X. Chen, "Personalized federated learning for intelligent IoT applications: A cloud-edge based framework," *IEEE Open J. Comput. Soc.*, vol. 1, pp. 35–44, 2020.

[27] Y. Laguel, K. Pillutla, J. Malick, and Z. Harchaoui, "Device heterogeneity in federated learning: A superquantile approach," 2020, *arXiv:2002.11223*.

[28] B. Li, W. Gao, J. Xie, M. Gong, L. Wang, and H. Li, "Prototype-based decentralized federated learning for the heterogeneous time-varying IoT systems," *IEEE Internet Things J.*, early access, Sep. 11, 2023, doi: 10.1109/JIOT.2023.3313118.

[29] Z. Ma, M. Xiao, Y. Xiao, Z. Pang, H. V. Poor, and B. Vucetic, "High-reliability and low-latency wireless communication for Internet of Things: Challenges, fundamentals, and enabling technologies," *IEEE Internet Things J.*, vol. 6, no. 5, pp. 7946–7970, Oct. 2019.

[30] Z. Ji, L. Chen, N. Zhao, Y. Chen, G. Wei, and F. R. Yu, "Computation offloading for edge-assisted federated learning," *IEEE Trans. Veh. Technol.*, vol. 70, no. 9, pp. 9330–9344, Sep. 2021.

[31] D. Ha, A. Dai, and Q. V. Le, "HyperNetworks," 2016, *arXiv:1609.09106*.

[32] A. Shamsian, A. Navon, E. Fetaya, and G. Chechik, "Personalized federated learning using hypernetworks," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 9489–9502.

[33] A. Z. Tan, H. Yu, L. Cui, and Q. Yang, "Towards personalized federated learning," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 34, no. 12, pp. 9587–9603, Dec. 2023.

[34] Q. Li, B. He, and D. Song, "Model-contrastive federated learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 10713–10722.

[35] Y. Huang, L. Chu, Z. Zhou, L. Wang, J. Liu, J. Pei, and Y. Zhang, "Personalized cross-silo federated learning on non-IID data," in *Proc. AAAI Conf. Artif. Intell.*, May 2021, vol. 35, no. 9, pp. 7865–7873.

[36] V. Smith, C.-K. Chiang, M. Sanjabi, and A. S. Talwalkar, "Federated multi-task learning," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 1–11.

[37] Y. Mansour, M. Mohri, J. Ro, and A. Theertha Suresh, "Three approaches for personalization with applications to federated learning," 2020, *arXiv:2002.10619*.

[38] Z. Yang, Y. Liu, S. Zhang, and K. Zhou, "Personalized federated learning with model interpolation among client clusters and its application in smart home," *World Wide Web*, vol. 26, no. 4, pp. 2175–2200, Jul. 2023.

[39] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," 2015, *arXiv:1503.02531*.

[40] D. Li and J. Wang, "FedMD: Heterogenous federated learning via model distillation," 2019, *arXiv:1910.03581*.

[41] P. Venkateswaran, V. Isahagian, V. Muthusamy, and N. Venkatasubramanian, "FedGen: Generalizable federated learning for sequential data," 2022, *arXiv:2211.01914*.
[42] T. Lin, L. Kong, S. U. Stich, and M. Jaggi, "Ensemble distillation for robust model fusion in federated learning," in *Proc. NIPS*, vol. 33, 2020, pp. 2351–2363.

SACHIN D. N received the bachelor's and master's degrees in computer science and engineering from Visvesvaraya Technological University (VTU), Belagavi, in 2013 and 2016, respectively. He is currently pursuing the Ph.D. degree with the Computer Science and Engineering Department, National Institute of Technology Karnataka (NITK), Surathkal, under the supervision of Prof. Annappa B. He is an Assistant Professor with the Vidya Vardhaka College of Engineering, India. His research interests include addressing data challenges, the heterogeneity of edge devices in federated learning, personalized federated learning, and privacy-preserving artificial intelligence.

ANNAPPA B (Senior Member, IEEE) received the B.E. degree from the University B.D.T. College of Engineering, Davangere, affiliated Mysore University, Karnataka, and the M.Tech. and Ph.D. degrees in computer science and engineering from the National Institute of Technology Karnataka, Surathkal, India. He is currently a Professor with the Department of Computer Science and Engineering, National Institute of Technology Karnataka, Surathkal. He has more than 25 years of experience in teaching and research. He has published more than 100 research papers in international conferences and journals. His research interests include cloud computing, big data analytics, distributed computing, software engineering, and process mining. He is a Life Member of the Computer Society of India, the Indian Society of Technical Education, the Cloud Computing Innovation Council of India, and the Advanced Computing and Communications Society. He is a fellow of the Institution of Engineers, India. He was the Organizing Chair of the International Conference ADCONS-2013, the Chair of the IEEE Mangalore Subsection, in 2018, and the General Chair of DISCOVER, in 2010. He serves on the TPC of many international conferences and is a reviewer for several journals. He is the Chair of the IEEE Computer Society Chapter, India Council. He was the Secretary of the IEI Mangaluru Local Center.

SAUMYA HEGDE (Senior Member, IEEE) received the M.Tech. and Ph.D. degrees from the National Institute of Technology Karnataka, Surathkal, India. She is currently an Assistant Professor with the Computer Science and Engineering Department, National Institute of Technology Karnataka, Surathkal. She has also conducted consultancy projects in domain name systems, particularly DoH and DoQ. Her research interest includes software-defined networking, with a focus on scalability issues.

CHUNDURU SRI ABHIJIT is currently pursuing the B.Tech. degree in computer science and engineering with the Vellore Institute of Technology. He is currently an experienced Research Intern with a strong background in federated learning, natural language processing, and ML, and has contributed to innovative projects in healthcare, privacy preservation, and smart grid optimization.

SATEESH AMBESANGE is currently a Research Scholar with NITK Surathkal, where he is dedicated to advancing academic knowledge and supporting the growth of the Indian startup ecosystem. His career is characterized by a strategic vision and a commitment to fostering innovation. In addition to his corporate accomplishments, he has more than two years of experience leading a startup, guiding it through accelerators, and securing angel funding. He brings over 15 years of experience in product development, showcasing leadership in initiating new projects, and generating innovative ideas for senior management. His expertise spans various product and team management roles, contributing significantly to the success of different technology products.