Ethical AI for Healthcare Systems: Uncertainty-Aware, Fair Federated Learning

Dian Chen, Qi Zhang
{dianc,qiz21}@vt.edu
Computer Science, Virginia Tech
USA

Lance Kaplan
lance.m.kaplan.civ@army.mil
US DEVCOM Army Research Laboratory
USA

Audun Jøsang
josang@ifi.uio.no
Informatics, University of Oslo
USA

Donghyun Jeong
djeong@udc.edu
CSIT, University of the District of Columbia
USA

Feng Chen
feng.chen@utdallas.edu
The University of Texas at Dallas
USA

Jin-Hee Cho
jicho@vt.edu
Computer Science, Virginia Tech
USA

ABSTRACT

This paper proposes U-FARE, an uncertainty-aware fair federated learning (FL) framework aimed at improving disease prediction in healthcare, with a specific focus on Alzheimer's disease detection. U-FARE incorporates evidential neural networks (ENN) to quantify uncertainty, enhancing both model fairness and accuracy. The framework ensures group-level fairness, providing consistent model performance across diverse healthcare environments despite data heterogeneity. We evaluate U-FARE on three real-world healthcare datasets—NACC, OASIS, and ADNI—comparing its performance to several state-of-the-art fairness-aware FL methods. Experimental results demonstrate that U-FARE outperforms baseline methods in both prediction accuracy and fairness, effectively balancing these two crucial aspects. The results also reveal the trade-off between fairness and accuracy, where higher fairness levels may compromise prediction accuracy. U-FARE achieves the highest accuracy (0.928) on the NACC dataset, consistently outperforms the competitive baseline q-FedAvg by 46%, particularly when higher fairness constraints are applied, and outperforms methods like Ditto and q-FFL with minimal accuracy variance and loss disparity. This is the first approach to simultaneously optimize fairness and accuracy in FL for Alzheimer's disease detection, providing a novel solution to the challenge of fair and effective AI in healthcare. The framework demonstrates the potential to address data heterogeneity while ensuring privacy and fairness in real-world applications.

CCS CONCEPTS

• **Computing methodologies** → *Cooperation and coordination*; **Distributed artificial intelligence**; **Reasoning about belief and knowledge**; **Machine learning**; • **Security and privacy** → *Privacy-preserving protocols*; • **Applied computing** → *Health care information systems.*

KEYWORDS

Ethical AI, healthcare systems, evidential uncertainty, federated learning, fairness, evidential neural networks

ACM Reference Format:
Dian Chen, Qi Zhang, Lance Kaplan, Audun Jøsang, Donghyun Jeong, Feng Chen, and Jin-Hee Cho. 2025. Ethical AI for Healthcare Systems: Uncertainty-Aware, Fair Federated Learning. In *ACM/IEEE International Conference on Connected Health: Applications, Systems and Engineering Technologies (CHASE '25), June 24–26, 2025, New York, NY, USA.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3721201.3721367

1 INTRODUCTION

The advent of artificial intelligence (AI) in healthcare has revolutionized disease detection, diagnosis, and management, particularly in complex conditions like Alzheimer's disease, where early diagnosis is crucial for timely intervention and improved quality of life [1, 34]. However, medical data heterogeneity – arising from differences in demographics, imaging modalities, and clinical protocols – makes it challenging to develop models that perform uniformly well across diverse populations [1, 34]. Federated learning (FL) has emerged as a promising approach for training models collaboratively while preserving data privacy, facilitated by the proliferation of distributed medical devices and patient data [30]. Yet, achieving ethical AI in healthcare through FL presents challenges, particularly in ensuring fairness and addressing uncertainty in predictions. Fairness is essential to avoid exacerbating disparities in healthcare access and outcomes. FL often focuses on group fairness, as data quality, quantity, and distribution vary across devices [7]. Without group fairness, disparities in model performance can lead to inequitable outcomes, undermining trust in AI systems [7, 30].

Uncertainty in healthcare AI is a critical factor influencing fairness, particularly in disease prediction, where it can affect patient outcomes [25, 36]. The heterogeneous and non-IID nature of data in FL can amplify uncertainty, causing uneven learning and unreliable predictions across the network, which impacts fairness and reliability [23]. Uncertainty-aware approaches offer several advantages in FL. They allow quality-aware aggregation by weighting

CHASE '25, June 24–26, 2025, New York, NY, USA

ACM ISBN 979-8-4007-1539-6/2025/06…$15.00
https://doi.org/10.1145/3721201.3721367

contributions based on reliability, preventing low-quality or outlier data from skewing the global model [32]. These models also support personalized adaptation, improving local performance while maintaining global fairness, and are robust to adversarial or noisy data [32]. Uncertainty estimates enhance trust and interpretability, allowing clinicians to make informed decisions [25]. One of the key technical challenges lies in integrating uncertainty into fairness to improve the performance of FL-based healthcare systems. Therefore, this work proposes an uncertainty-aware fair FL framework that ensures high prediction accuracy in Alzheimer's detection while maintaining robust group fairness, even with data heterogeneity. We make the following **key contributions**:

- We propose an uncertainty-aware, fair federated learning (FL) framework, named U-FARE: Uncertainty-aware, FAir fedeRated lEarning, designed explicitly for healthcare systems in disease prediction. The framework utilizes evidential neural networks (ENN) to quantify uncertainty during the learning process, enhancing model performance (i.e., prediction accuracy) and fairness. This is the first approach to integrate ENNs into FL settings.
- We ensure *group-level fairness* in the uncertainty-aware FL framework, guaranteeing consistent model performance across all clients. Group-level fairness ensures that the model performs equitably across different groups of clients, even when their data varies in quality, quantity, or distribution [11]. By addressing disparities in data quality and distribution, the framework promotes equitable outcomes, ensuring that each client, regardless of their data characteristics, benefits from uniform model performance in diverse healthcare settings. Furthermore, unlike existing works, we explicitly consider fairness threats in testing our fairness-aware FL framework. This provides a more rigorous evaluation of fairness mechanisms under adversarial conditions, ensuring robustness in real-world applications.
- We evaluate the proposed approach on real-world healthcare datasets, marking the first application of uncertainty-aware fair FL to Alzheimer's disease detection [17, 27]. The results demonstrate the superior performance of U-FARE, highlighting its effectiveness in balancing fairness and prediction accuracy while achieving state-of-the-art results. To date, no prior work has addressed the challenge of optimizing both fairness and model accuracy using FL for Alzheimer's disease detection.
- The framework provides a comprehensive comparison with existing fairness-aware FL methods, demonstrating U-FARE's superior performance in terms of prediction accuracy, fairness, accuracy variance (AV), and loss disparity (LD). This benchmarking showcases U-FARE's robustness in handling data heterogeneity while maintaining fairness.
- We present a novel sensitivity analysis on varying fairness levels, illustrating the trade-off between fairness and prediction accuracy. This analysis provides key insights into how the system balances these metrics effectively and guides the fine-tuning of the framework to achieve the desired outcome.

2 RELATED WORK

2.1 Fairness-aware Federated Learning

Mohri et al. [29] introduced Agnostic Federated Learning (AFL), optimizing model performance across all client distributions by minimizing worst-case loss, ensuring robustness and fairness in heterogeneous settings. Li et al. [22] proposed q-Fair Federated Learning (q-FFL), an optimization framework for fairness in federated systems with non-IID data, prioritizing clients with higher loss values to reduce performance disparities. q-FFL includes three algorithms: q-FedAvg, which extends Federated Averaging with client loss-based weighting; q-FedSGD, which adapts stochastic gradient descent; and q-MAML, integrating fairness into Model-Agnostic Meta-Learning (MAML) for better client task generalization.

Li et al. [21] developed Ditto, a personalized federated learning framework optimizing both global and individual client models, allowing adaptation to unique data while benefiting from shared knowledge. However, Ditto lacks specific fairness metrics for evaluation. Liu et al. [24] proposed the Contribution-Aware Federated Learning (CAreFL) framework for smart healthcare, improving model aggregation efficiency by 2.84 times while ensuring fair, explainable, and privacy-preserving evaluations. Düsing and Cimiano [10] introduced the Benefit and Contribution metrics, showing that data imbalances reduce benefits while increasing client contributions. Hosseini et al. [13] developed Proportionally Fair Federated Learning (Prop-FFL) to reduce performance variations across hospitals, while Cui et al. [8] used multi-objective optimization to ensure fairness and performance consistency across local clients.

Ezzeldin et al. [11] proposed FairFed, a fairness-aware aggregation algorithm to enhance group fairness in FL by evaluating fairness on local datasets and adjusting aggregation weights by aligning global fairness metrics, which ensures privacy via secure aggregation. Li et al. [23] adapted the Gini coefficient to quantify fairness in FL, measuring fairness in model accuracy across clients.

While most works focus on collaborative fairness, they often neglect group fairness, especially in healthcare settings. Li et al. [22] addresses group fairness but does not apply it to real-world datasets like healthcare data. Our approach addresses these gaps by introducing an uncertainty-aware fair method tailored for healthcare datasets, specifically Alzheimer's disease, while enhancing group fairness and overcoming the limitations of prior work.

2.2 FL-based Disease Diagnosis and Detection

Meerza et al. [27] introduced the first FL approach for automatic Alzheimer's Disease (AD) diagnosis using spontaneous speech analysis, ensuring fairness across clients. They employed q-FEDAvg and q-FEDSgd aggregation mechanisms to mitigate algorithmic bias from data heterogeneity. Zhang et al. [42] applied cryptographic techniques in IoT and FL-based healthcare systems to protect local models from adversarial attacks like model reconstruction. Yazdinejad et al. [41] proposed the AP2FL model, using Trusted Execution Environments (TEE) to secure both clients and servers during training, while integrating Active Personalized Federated Learning (ActPerFL) and Batch Normalization (BN) to mitigate performance degradation due to data heterogeneity.

Li et al. [20] developed ADDetector, an FL-based AD detection system that ensures privacy against man-in-the-middle attacks through asynchronous aggregation and enhances accuracy with linguistic features from smart speakers. Mitrovska et al. [28] compared FedAvg and secure aggregation for AD detection under data heterogeneity and member inference attacks. Khalil et al. [17] deployed an FL-based method to create a shared AD prediction model without accessing sensitive local data, improving training efficiency via hardware acceleration. Kumar et al. [18] proposed a blockchain-based FL model for COVID-19 detection using CT imaging and introduced data normalization to address data heterogeneity. Similarly, Singh et al. [37] implemented an FL and blockchain framework for patient monitoring in smart healthcare.

While some works address fairness, many lack comprehensive evaluation strategies and fairness metrics. Moreover, few consider fairness attacks, which are essential for evaluating fairness under adversarial conditions. This study examines adversarial attacks on FL systems and employs fairness metrics to assess our approach using real-world healthcare datasets.

2.3 Uncertainty-Aware AI-based Disease Diagnosis and Detection

MacDonald et al. [25] emphasized the role of uncertainty in AI healthcare, particularly in decision-making, safety, and reliability. It helps quantify confidence in predictions, is critical for high-risk domains like clinical decision-making, and supports fairness by identifying model bias, especially in underrepresented populations. Tabarisaadi et al. [38] improved AI reliability in skin cancer detection by integrating uncertainty quantification (UQ) into models, using algorithms like MC Dropout, Bayesian Ensembling, and SNGP. Ghoshal et al. [12] applied Bayesian Convolutional Neural Networks (BCNN) for grading pancreatic adenocarcinoma, using predictive uncertainty to flag unreliable predictions. Prince et al. [31] introduced Bayesian deep learning to enhance the diagnosis of adamantinomatous craniopharyngioma (ACP) from MRI images, identifying uncertainty through predictive distributions. Wang et al. [40] proposed FedUAA, an FL model for diabetic retinopathy, incorporating uncertainty-aware weighting to evaluate client reliability and adapt aggregation weights.

While uncertainty-aware models improve reliability, their calibration often lacks consistency across demographic subgroups, leading to varying confidence levels. Existing works overlook security challenges, such as adversarial attacks, which degrade model performance. In contrast, our approach uses uncertainty-aware techniques to enhance both model performance and fairness in FL healthcare systems, even under adversarial conditions.

3 PRELIMINARIES

3.1 Evidential Neural Networks (ENNs)

An Evidential Neural Network (ENN) is similar to classical neural networks, with the key difference being that the softmax layer is replaced by an activation layer (i.e. ReLU) to ensure non-negative outputs. These outputs are then treated as the evidence vector, which is used to model a predicted Dirichlet distribution [35] built on the theory of *Dempster-Shafer theory* [9] and *Subjective Logic* [15]. Instead of outputting a single class label, it produces a belief function that quantifies the confidence in various possible outcomes. Hu et al. [14] further enhanced [35]'s work, allowing it to measure the quality of the predicted Dirichlet distributions directly. For a given sample i, let $f(\mathbf{x}_i|\Theta)$ denote the evidence vector predicted by the network for classification, where $\mathbf{x}_i \in \mathbb{R}^L$ represents the input features and Θ corresponds to the network parameters. The associated Dirichlet distribution has parameters $\boldsymbol{\alpha}_i = f(\mathbf{x}_i|\Theta) + 1$. Let y_i represent the true label. The Dirichlet density $\mathrm{Dir}(\mathbf{p}_i; \boldsymbol{\alpha})$ serves as the prior for the Multinomial distribution, $\mathrm{Multi}(y_i|\mathbf{p}_i)$. To estimate the parameters $\boldsymbol{\alpha}_i$ for sample i, the following sum of squared loss is formulated, given the Dirichlet PDF:

$$\begin{aligned}\mathcal{L}(f(\mathbf{x}_i|\Theta), y_i) &= \int \frac{\|y_i - \mathbf{p}_i\|_2^2}{B(\boldsymbol{\alpha}_i)} \prod_{j=1}^{K} p_{p_{ij}}^{(\alpha_{ij}-1)} \, d\mathbf{p}_i \qquad (1)\\ &\sim \sum_{j=1}^{K} \left(y_{ij}^2 - 2y_{ij}\mathbb{E}[p_{ij}] + \mathbb{E}[p_{ij}^2] \right),\end{aligned}$$

where $B(\boldsymbol{\alpha}) = \frac{\prod_{i=1}^{K} \Gamma(\alpha_i)}{\Gamma\left(\sum_{i=1}^{K} \alpha_i\right)}$ and $\mathbb{E}$ is the expected probability distribution. The k-dimensional Dirichlet PDF represents the multinomial probability density over a domain with cardinality k, which can reduce to a Beta PDF with the specific case of a binary domain ($k = 2$) as a binary opinion. This work uses the Beta PDF to represent a binomial opinion in binary decision-making for disease prediction.

3.2 Problem Statement

This work uses an uncertainty-aware FL framework to detect Alzheimer's disease while ensuring group-level fairness. The system is structured as a horizontal FL model, with a global model hosted on a cloud server and local models operating across hospitals (clients). This decentralized approach maintains data privacy, as each hospital trains its local model using patient data and sends periodic updates to the global model. The challenge is ensuring fair updates from local models, handling data heterogeneity across hospitals.

This work aims to propose a solution to these challenges by introducing U-FARE (Uncertainty-aware, FAir fedeRated lEarning), a novel uncertainty-aware, fair FL framework designed specifically for healthcare disease prediction tasks. U-FARE leverages evidential neural networks (ENN) to have the model performance of each client uniformly distributed (group fairness) while maintaining the high prediction accuracy of the predictive model. We formulate the objective by:

$$\text{maximize} \quad GF(M(\theta^*)), \quad \text{subject to } \mathcal{ACC}(M(\theta^*)) \geq \tau. \qquad (2)$$

In our system, $M(\theta^*)$ denotes the testing prediction accuracy across the local models with parameters θ^* after achieving the total communication rounds. θ^* is a set of local model parameters each client learns. $GF(M(\theta^*))$ is group fairness, which will be measured by specific fairness metrics (see Section 6.4), and τ is a threshold for the testing accuracy. This approach underlines our commitment to developing an uncertainty-aware, fair FL for the healthcare system, achieving high prediction accuracy while guaranteeing group fairness across participants.

4 SYSTEM MODEL

4.1 Network Model

The smart healthcare system considered here employs a network model that integrates multiple hospitals as clients and a cloud server as the central server, as illustrated in Figure 1. This FL structure is

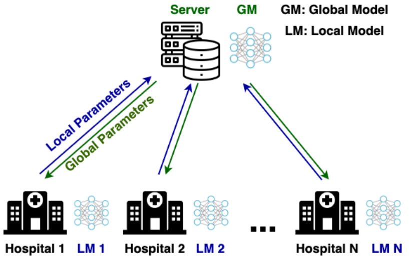


Figure 1: Network model of FL healthcare system.

known as horizontal federated learning (HFL), where data share the same feature space but come from different sample spaces. In an FL-based smart healthcare system, each hospital collects individual data, such as medical images (feature space), from multiple patients (sample space) to train its local predictive model. While hospitals may treat different patients, they all use the same types of information, such as MRI images, to detect specific diseases. For FL operations, the system begins with the central server initializing a global model, which is then distributed to each client (i.e., hospital). Each client trains the model locally using its patient data, computes updates, and sends these updates back to the central server. The server aggregates these updates, refines the global model, and redistributes the improved model to all clients. This iterative process of local training, update aggregation, and global model refinement continues for several rounds until the model reaches satisfactory performance. A deep learning (DL) model, deployed across clients and the central server, is used to detect Alzheimer's disease, outputting a binary classification: 0 for healthy and 1 for disease diagnosis. This approach enables the development of robust predictive models by leveraging diverse datasets from multiple hospitals while ensuring patient data privacy and security.

4.2 Node Model

In the healthcare network, the clients are individual hospitals. Each hospital possesses its own local dataset, which includes patient information, medical records, imaging data, etc. The hospitals do not share their raw data with other institutions to maintain patient privacy and comply with data protection regulations. Each hospital trains a local model on its dataset. During this process, the hospital's data never leaves its premises. Instead, the hospital computes updates to the model (such as gradients) based on its local data. These updates are then sent to the central server. This setup allows hospitals to collaborate on developing robust predictive models without compromising patient confidentiality. It is challenging to create a more diverse and generalized model since the data from different hospitals can vary significantly in terms of demographics, disease prevalence, and medical practices.

The central server acts as the aggregator and coordinator of the federated learning process. It does not have direct access to any raw data from the hospitals. The central server collects model updates (e.g., gradients or model parameters) from all participating clients. It aggregates these updates to create a global model. This global model is then sent back to the clients, which further train it using their local data in the next iteration. This process is repeated for multiple rounds until the model converges to a satisfactory performance level. The central server facilitates collaboration among hospitals, ensuring that the benefits of shared learning are realized without compromising data privacy. It orchestrates the overall learning process, ensuring that the global model continually improves by leveraging the diverse datasets of all participating hospitals.

4.3 Threat Model

To understand the vulnerabilities within FL systems and evaluate the effectiveness of our approach against different attacks, we examine adversarial attacks performed on compromised clients aiming to disrupt the training process by modifying the local model parameters or local training datasets. We consider them *fairness attacks* since malicious updates can further create performance discrepancies among clients, violating the fairness principle of uniform local performance [33].

- ***Byzantine attacks*** [5] primarily target clients (i.e., hospitals), prolonging their learning duration or leading to model divergence while training their local models. They disrupt the FL training process by injecting arbitrary metrics via Stochastic Gradient Descent (SGD) updates. In this work, we employ a Gaussian noise-based approach to introduce Gaussian noise into the model parameters of compromised clients before they are uploaded to the central server. The success of Byzantine attacks introduces Byzantine updates that can skew the global model to perform poorly on specific data distributions, affecting only a subset of participants. Clients with local data distributions closer to the poisoned updates are disproportionately affected, creating unequal performance across participants.
- ***Poisoning attacks*** [39] compromise the datasets on clients that deliberately manipulate input data to deceive the model into making incorrect predictions. These manipulations are often minimal and invisible to humans but can cause significant errors in the model's output. This work considers a backdoor-based approach [3], in which a malicious client injects a specific pattern or trigger into its local training data to manipulate the global model's behavior. The attacker aims to make the model perform generally on standard inputs but produce attacker-desired outputs when the trigger is present. Such an attack can cause the global model to learn biased or incorrect patterns, disproportionately affecting local devices with data that overlaps with poisoned classes. For instance, if certain classes are targeted for backdooring, participants with a higher proportion of these classes in their local datasets will experience degraded performance, breaking the fairness guarantee.

5 PROPOSED APPROACH: U-FARE

In the FL context, group fairness differs from its interpretation in traditional deep learning domains. A group-level fair solution means more *uniform* regarding various metrics. One common metric for uniformity is the variance of accuracy distribution, which we can formally define the notion of fairness [22].

Definition 1 [22]: A solution w is more fair than $w^{`}$ if the performance of m devices $\{F_1, ..., F_m\}$ satisfies

$$\mathbf{Var}(F_1(w), \ldots, F_m(w)) < \mathbf{Var}(F_1(w^{'}), \ldots, F_m(w^{'})). \qquad (3)$$

Ensuring fairness in a smart healthcare system for disease detection is crucial for equitable healthcare outcomes across hospitals. Disparities in model performance can lead to unequal access to accurate diagnoses, exacerbating healthcare inequalities. By ensuring consistent model performance across hospitals, we promote reliable healthcare services regardless of geographic or demographic differences. This enhances trust in the system and ensures high-quality patient care, improving overall public health outcomes. Therefore, developing a fair FL framework is essential for building a robust and adaptable healthcare system that serves diverse populations.

To achieve group-level fairness in an ENN-based FL framework, we propose an algorithm integrating the uncertainty and fairness by introducing a concept of the *degree of conflict*. As described before, An ENN model is built on *Subjective Logic* (SL), which can formulate a belief model as the output of the ENN model to explicitly deal with uncertainty. In SL [15], a client A can form its opinion about a given proposition X, denoted by $\omega_X^A = \{\boldsymbol{b}_X, u_X, \boldsymbol{a}_X\}$, where $\boldsymbol{b}_X$ is the belief masses distribution, u_X is the uncertainty mass, and $\boldsymbol{a}_X$ is the base rate (i.e., prior belief) distribution of variable X. The components satisfy the additivity requirement with $u_X + \sum \boldsymbol{b}_X(x) = 1$. A core assumption of SL is that different agents can hold varying opinions about the same variable. This demonstrates the subjective way people perceive the world.

Regarding predictive tasks, having different opinions about the same sample can be a significant issue since it complicates determining the best prediction outcome. Accordingly, the *degree of conflict* (DC) [16] measures the disparity between opinions and can be utilized to manage differing views about the same target. Consider two agents, B and C, each holding opinions ω_X^B and ω_X^C about a common variable X. A fundamental way to quantify the conflict between these opinions is through the *projected distance* (PD), as:

$$PD(\omega_X^B, \omega_X^C) = \frac{1}{2} \sum_{x \in X} \left| P_X^B(x) - P_X^C(x) \right|. \tag{4}$$

The PD has the property $PD \in [0, 1]$. When $PD = 0$, it indicates identical projected probability distributions, signifying no conflict, though the underlying opinions may differ. Conversely, $PD = 1$ corresponds to maximum disagreement, arising when the projected probabilities represent completely opposing views. However, a high PD does not always show conflict, as high uncertainty in one or both opinions can mitigate the potential disagreement. When uncertainty is high, a greater PD can be tolerated since uncertain opinions contribute less weight in any fusion process.

The *conjunctive certainty* (CC) [15] is then leveraged as a logical metric to quantify the combined certainty of two opinions, ω_X^B, and ω_X^C, and is defined as:

$$CC(\omega_X^B, \omega_X^C) = (1 - u_X^B)(1 - u_X^C). \tag{5}$$

The value of CC lies in the range $[0, 1]$, where $CC = 0$ indicates complete uncertainty in at least one opinion, and $CC = 1$ means both opinions are fully certain with no uncertainty.

The *degree of conflict* (DC) between the opinions is defined as the product of PD and CC.

Definition 2 [15]: Consider two agents, B and C, who hold respective opinions ω_X^B and ω_X^C regarding a shared variable X. The degree of conflict (DC) between these opinions, denoted as $DC(\omega_X^B, \omega_X^C)$, is defined as follows:

$$DC(\omega_X^B, \omega_X^C) = PD(\omega_X^B, \omega_X^C) \cdot CC(\omega_X^B, \omega_X^C). \tag{6}$$

With our FL framework, client A's current opinion about a patient is significantly in conflict with the updated global model's opinion ω_g, and client B's opinion has low DC with ω_g about the same patient, the prediction accuracy of client A and B will be different and result in a large variance of the performance distribution which means unfair. Therefore, we leverage the concept of DC to achieve fairness in clients by reducing *degree of conflict* between clients' opinions and the central server's opinion.

Our framework follows the training procedure of FedAvg [26] and extends with the proposed uncertainty-aware approach. At the beginning of each communication round t, the central server randomly selects K clients to participate in the FL training process. The server then transmits the global model parameters θ_t to the selected clients. Each client k updates θ_t for E epochs using its local dataset X_k, resulting in the updated model θ_t^k.

With the updated local model, client k computes the parameters of the Dirichlet distribution as follows:

$$\boldsymbol{\alpha}_{X_k}^k = \{\boldsymbol{\alpha}_i \mid i \in X_k\} = \{f(x_i|\theta_t^k) + 1 \mid i \in X_k\}, \tag{7}$$

where $f(x_i|\theta_t^k)$ represents the evidence vector predicted by the network for classification tasks on client k's local dataset, and $x_i \in X_k$ denotes the input feature.

Next, given W is the number of classes in X_k, client k calculates the belief mass $\boldsymbol{b}_{X_k}^k$ and uncertainty $u_{X_k}^k$ by:

$$\boldsymbol{b}_{X_k}^k = \frac{f(x_i|\theta_t^k)}{S}, \quad u_{X_k}^k = \frac{W}{S}, \tag{8}$$

where $S = \sum_{x_i \in X_k} (f(x_i|\theta_t^k) + 1)$.

Finally, each client k sends its updated model θ_t^k and its opinion, represented as $\omega_{X_k}^k = (\boldsymbol{b}_{X_k}^k, u_{X_k}^k, \boldsymbol{\alpha}_{X_k}^k)$, back to the central server. **Algorithm 1** demonstrates the previous steps with lines 1-6.

After receiving the local updates from the clients, the central server updates the global model parameters θ_{t+1} using the standard FedAvg approach:

$$\theta_{t+1} = \frac{\sum_{i=1}^{K} \theta_i^t}{K}, \tag{9}$$

where K is the total number of selected clients, the server computes the global opinion ω_X^g based on θ_{t+1}, and X represents the global dataset. Following this, the central server evaluates the DC between the global opinion ω_X^g and each selected client's opinion $\omega_{X_k}^k$, based on Eq. (6). The server sorts the selected clients in descending order based on their DC with the global opinion to identify the client with the highest conflict with the global model. The server recalculates the global model by assigning a higher weight to the most conflicting client, scaled by a factor λ. This highlights the primary innovation of the proposed algorithm, which leverages uncertainty quantification, distinguishing it from traditional fairness-aware approaches, such as q-FFLs [22], which adjust a client's influence on the global model based on its loss value. **Algorithm 1** describes this process in lines 7-10.

Enhancing the influence of the client that most conflicts with the global model can help reduce performance disparities but may also compromise the prediction accuracy for that client. To address

Algorithm 1 U-FARE

```
Require: Total number of clients $N$, number of selected clients $K$,
    clients' datasets $\{X_1, \dots, X_m\}$, total communication rounds $T$,
    Epoch $E$, initialized model parameters $\theta_0$, clients $k = 1, \dots, m$
for $t = 0, \dots, T-1$ do
    Central server randomly selects $K$ clients
    Central server sends local model $\theta_t$ to all selected clients
    Each selected client $k$ updates $\theta_t$ for $E$ epochs to obtain $\theta^k_t$
    Each client $k$ formulates its opinion $\omega^k_{X_k} = (\boldsymbol{b}^k_{X_k}, u^k_{X_k}, \boldsymbol{\alpha}^k_{X_k})$
    Each client $k$ sends $\theta^k_t$ and opinion $\omega^k_{X_k}$ to central server
    Central server updates $\theta_{t+1}$
    Central server obtains global opinion $\omega^g_X$ based on $\theta_{t+1}$
    Find $\theta^* = \arg\max_{k=1}^{K} DC(\theta^g_X, \theta^k_{X_k})$
    Compute new $\theta_{t+1}$ by
        $$\theta_{t+1} = \frac{\sum_{k=1}^{K} \theta^t_k + \lambda\theta^*}{K},$$
    where $\lambda$ is the weight assigned to the client with the highest
    conflict.
end for
```

this, it is essential to determine an optimal scaling factor, λ, that improves the uniformity of local model performance while maintaining high overall prediction accuracy. We carefully fine-tune λ through empirical experiments, selecting the value that outperforms baseline approaches in terms of both fairness and prediction accuracy.

6 EXPERIMENTAL SETUP

6.1 Datasets

To evaluate our proposed approach, we utilize a collection of datasets specifically focused on Alzheimer's disease detection. Below, we provide a detailed description of each dataset and its relevance to Alzheimer's disease detection:

- **Alzheimer's Disease Neuroimaging Initiative (ADNI)** [2]: ADNI is a large, longitudinal, multi-site study launched in 2004 to advance research on Alzheimer's disease (AD). ADNI is one of the most significant and influential research projects in the field of neurodegenerative diseases. This dataset contains 3D volumetric magnetic resonance imaging (MRI) scans for binary classification of Alzheimer's disease vs. cognitive normal. They were preprocessed to extract measurements of regional volumes derived from neuroimaging data. The final dataset is tabular, consisting of 24,159 samples, each characterized by 87 features.
- **OASIS-3** [19]: The OASIS-3 dataset is a comprehensive retrospective resource for studying normal aging and Alzheimer's disease. It spans 30 years and includes data from 1,378 participants aged 42–95 years. This dataset includes MRI scans designed for binary classification between Alzheimer's disease and cognitively normal states, with a total of 40332 samples.
- **National Alzheimer's Coordinating Center (NACC)**: This database collects data through the Uniform Data Set (UDS), which has been collecting longitudinal standardized clinical data since 2005. Data are provided by Alzheimer's Disease Research Centers (ADRCs) as part of the National Institute of Aging's ADRC Program. We select an MRI dataset for binary classification of Alzheimer's disease vs. cognitive normal, including 16200 image samples.

Table 1: Key Design Parameters and Default Values

Parameter	Value
Total number of clients	5
Communication rounds	20
Local training epochs	40
Simulation runs	10
Learning rate	0.1
Learning rate lambda	0.1
Batch size	64

For each dataset, we employ the Synthetic Minority Over-sampling Technique (SMOTE) [6] to mitigate the class imbalance, ensuring an equal number of samples in both the positive and negative classes. Aside from SMOTE, no additional data preprocessing techniques are applied, as the primary objective of this work extends beyond merely improving prediction accuracy.

6.2 Parameterization

The hyperparameters used in the experiments are meticulously selected to optimize the performance of the federated learning framework. The learning rate is set to 0.1, along with a learning rate lambda of 0.1 to effectively balance the loss components. Each client performs local training using a batch size of 64 over 40 epochs per round. The framework comprises 20 communication rounds, with five clients actively participating in each round, and the simulations are repeated 10 times to ensure robust evaluation. A convolutional neural network (CNN) is utilized as the model, chosen for its efficiency and effectiveness in image classification. Table 1 provides a detailed summary of the key parameters and their default values used in the experiments.

6.3 Comparing Schemes

To evaluate the effectiveness of our proposed approach, we compare U-FARE with the following state-of-the-art fairness-aware schemes:

- **Agnostic FL (AFL)** [29] is a framework that optimizes a centralized model to perform well across any target distribution, formed by a mixture of client distributions, addressing biases in FL. This approach naturally promotes fairness and provides data-dependent guarantees for learning, along with a fast optimization algorithm and convergence bounds.
- **q-FedSGD** [22] is an optimization method designed to solve the q-Fair Federated Learning (q-FFL) problem using Stochastic Gradient Descent (SGD) in FL settings. It extends the idea of fairness in federated learning by adjusting the gradients to promote a more uniform accuracy distribution across participating devices.
- **q-FedAvg** [22] is a communication-efficient version of q-FedSGD, replacing the update steps with a heuristic approach.
- **q-MAML** [22] is a meta-learning method that extends the q-FFL objective to improve fairness in personalized models. It learns a model initialization that can be quickly adapted to new tasks with limited data, while reducing accuracy variance across tasks.

- **Ditto** [21] is a personalized federated learning framework that balances fairness and robustness in statistically heterogeneous networks by addressing competing data constraints and model poisoning attacks and performance uniformity across devices.

6.4 Metrics

We evaluate the performance of the proposed FL system by measuring prediction accuracy and (group) fairness through three metrics:

- **Prediction Accuracy** measures the consistency between the model's predicted results for detecting Alzheimer's disease and the actual observed values. We evaluate the model performance by capturing the testing accuracy of aggregation for each communication round.
- **Loss Disparity (LD)** [4] computes the variance in testing loss values across clients, formulated by:

$$LD = \frac{1}{N}\sum_{i=1}^{N}\mathrm{Var}(\mathcal{L}_i), \tag{10}$$

where $\mathcal{L}_i$ is the testing loss value of client i, and N is the total number of clients. A lower LD indicates a more equitable distribution of loss, and thus a more fair FL system.
- **Accuracy Disparity (AD)** [22] measures the variance in accuracy across clients, defined as:

$$\mathrm{AD} = \frac{1}{N}\sum_{i=1}^{N}\left(\mathrm{Accuracy}_i - \overline{\mathrm{Accuracy}}\right)^2, \tag{11}$$

where $\mathrm{Accuracy}_i$ is client i's prediction accuracy for detecting Alzheimer's disease, and $\overline{\mathrm{Accuracy}}$ is the average accuracy across clients. A lower AD represents a higher group fairness of predictive performance in the FL framework.

7 RESULTS & ANALYSIS

7.1 Performance Comparison Analyses

Table 2 presents a detailed performance comparison of various fairness-aware FL methods, and the proposed U-FARE on three healthcare-related datasets: NACC, OASIS, and ADNI. The table provides results for both IID (Independent and Identically Distributed) and non-IID (non-Independent and Identically Distributed) data distributions, which are typical scenarios in federated learning for healthcare. In terms of prediction accuracy (Acc), U-FARE consistently shows superior performance across all datasets and data distributions. For example, on the NACC dataset under IID conditions, U-FARE achieves the highest accuracy (0.928) compared to other methods, which demonstrates its effectiveness in making accurate predictions.

Furthermore, U-FARE stands out not only in terms of accuracy but also in its fairness metrics, which are represented by accuracy variance (AV) and loss disparity (LD). Both AV and LD are critical for evaluating the fairness of FL methods, with lower values being preferable. U-FARE consistently maintains the lowest AV and LD values, indicating its ability to reduce fairness issues such as disparate performance across different clients. For instance, on the NACC dataset under IID conditions, U-FARE has an AV of 0.0007 and LD of 0.0006, which is substantially better than other methods, including Ditto, which has a higher AV and LD. Under the non-IID conditions, where the data is more challenging due to heterogeneous distributions across clients, U-FARE continues to demonstrate its superiority in fairness. The method significantly reduces AV and LD while maintaining competitive accuracy. For example, in the ADNI dataset under non-IID conditions, U-FARE achieves an accuracy of 0.802, along with AV and LD values of 0.039 and 0.0497, respectively, outperforming the other methods in terms of both fairness and prediction accuracy.

Additionally, we observe an interesting inconsistency between the two fairness metrics. For instance, in the case of the NACC dataset, U-FARE outperforms all baseline methods in terms of prediction accuracy and accuracy variance, yet it falls short of q-FedAvg in terms of loss disparity. This discrepancy may arise because q-FedAvg is a loss-based approach that achieves fairness by assigning higher weights to clients with higher loss values during training. In contrast, U-FARE leverages an uncertainty-aware fair approach, where uncertainty more accurately represents prediction outcomes than loss values. This may explain why U-FARE achieves superior performance in balancing both accuracy and fairness, as it utilizes uncertainty to enhance the model's understanding of its predictions, leading to better overall outcomes.

7.2 Sensitivity Analyses

7.2.1 Varying the level of fairness.

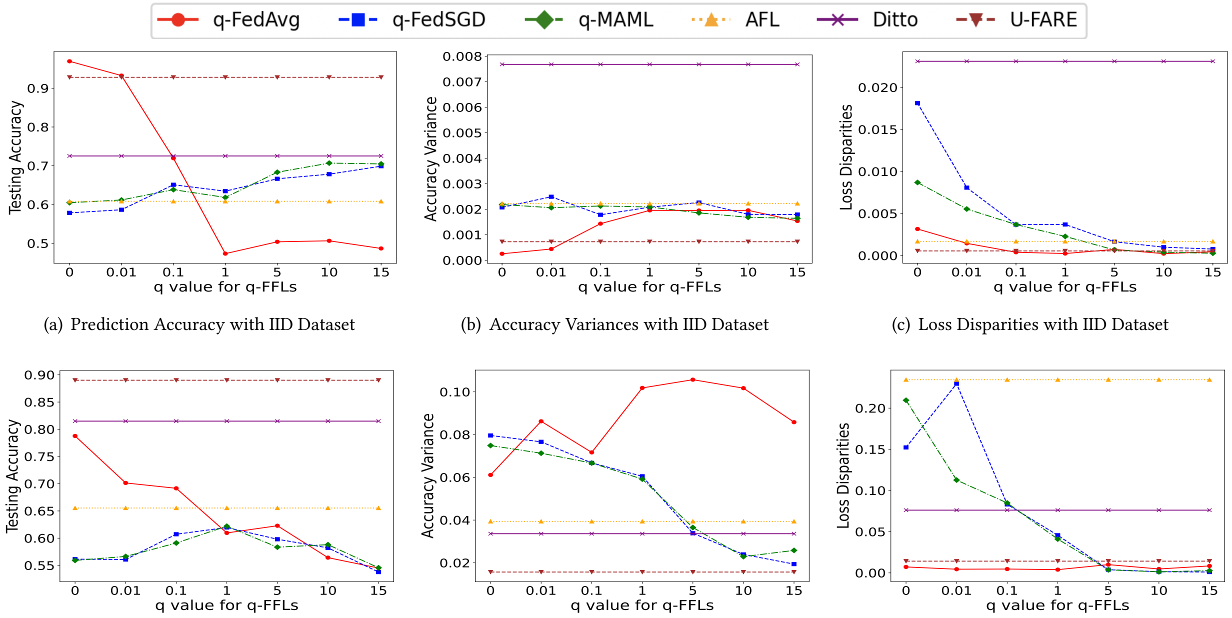
In our approach, fairness is not explicitly set or controlled, unlike in q-FFL methods, where fairness levels can be directly adjusted. To evaluate the performance of our proposed U-FARE, we conduct a sensitivity analysis by varying the fairness levels of the q-value-based baseline methods. This enables us to identify the specific fairness levels at which our approach outperforms the baseline methods, offering insights into its effectiveness under different fairness constraints.

Figures 2 to 4 show the effect of varying fairness levels, represented by the q value. The proposed U-FARE consistently outperforms all baseline methods across the three datasets, even at the highest fairness levels, except for q-FedAvg. For q-FedAvg, U-FARE performs better when $q \geq 0.01$ for IID data and consistently excels with non-IID data on the NACC dataset, as shown in Figure 2. On the OASIS dataset (Figure 3), U-FARE not only surpasses q-FedAvg in prediction accuracy but also achieves superior fairness for $q \geq 0.1$ with IID data, maintaining this advantage for all fairness levels with non-IID data. With the ADNI dataset in Figure 4, U-FARE outperforms q-FedAvg on both fairness metrics and achieves higher prediction accuracy for $q \geq 0.1$.

As expected, higher fairness levels often compromise prediction accuracy. An increase in q results in greater fairness, as measured by LD, but also a decrease in prediction accuracy due to the inclusion of lower-performing clients, which lowers overall model performance. However, this trend does not hold for the AV metric, which increases as q rises, indicating lower fairness. Although Li et al. [22] report different findings, our sensitivity analysis aligns with theirs in terms of performance comparison, suggesting that loss is not a robust metric for representing prediction outcomes or addressing fairness. Notably, in q-FFLs, the models and datasets used differ from those in this study, which could explain the discrepancies. This indicates that the q value may need fine-tuning for each specific setting (i.e.,

Table 2: Summary of Performance Comparison

Dataset	**Data Type**	**q-FedAvg** (q = 1)			**q-FedSGD** (q = 1)			**q-MAML** (q = 1)			**AFL**			**Ditto**			**U-FARE**		
		Acc	**AV**	**LD**	**Acc**	**AV**	**LD**	**Acc**	**AV**	**LD**	**Acc**	**AV**	**LD**	**Acc**	**AV**	**LD**	**Acc**	**AV**	**LD**
NACC	IID	0.473	0.0020	**0.0002**	0.635	0.0021	0.0037	0.618	0.0021	0.0023	0.602	0.0021	0.0017	0.736	0.0073	0.0231	**0.928**	**0.0007**	0.0006
	Non-IID	0.610	0.1018	**0.0039**	0.620	0.0605	0.0457	0.622	0.0592	0.0410	0.639	0.0487	0.2345	0.752	0.0566	0.0762	**0.890**	**0.0157**	0.0142
OASIS	IID	0.499	0.0016	**0.0006**	0.603	0.0015	0.0015	0.593	0.0015	0.0013	0.589	0.0015	0.0007	0.600	0.0055	0.0026	**0.832**	**0.0012**	0.0008
	Non-IID	0.493	0.0706	0.0102	0.554	0.0597	0.0378	0.573	0.0495	0.0296	0.563	0.0444	0.1993	0.680	0.0383	0.0214	**0.837**	**0.0077**	**0.0018**
ADNI	IID	0.679	0.0025	0.0006	0.838	0.0014	0.0007	0.858	0.0015	0.0006	0.817	0.0017	**0.0001**	0.818	0.019	0.0376	**0.912**	**0.0010**	0.0015
	Non-IID	0.652	0.0220	0.0096	0.799	0.0054	0.0031	0.801	0.0057	0.0028	0.817	0.0039	**0.0004**	0.802	0.039	0.0497	**0.921**	**0.0044**	0.0037



(a) Prediction Accuracy with IID Dataset

(b) Accuracy Variances with IID Dataset

(c) Loss Disparities with IID Dataset

(d) Prediction Accuracy with Non-IID Dataset

(e) Accuracy Variances with Non-IID Dataset

(f) Loss Disparities with Non-IID Dataset

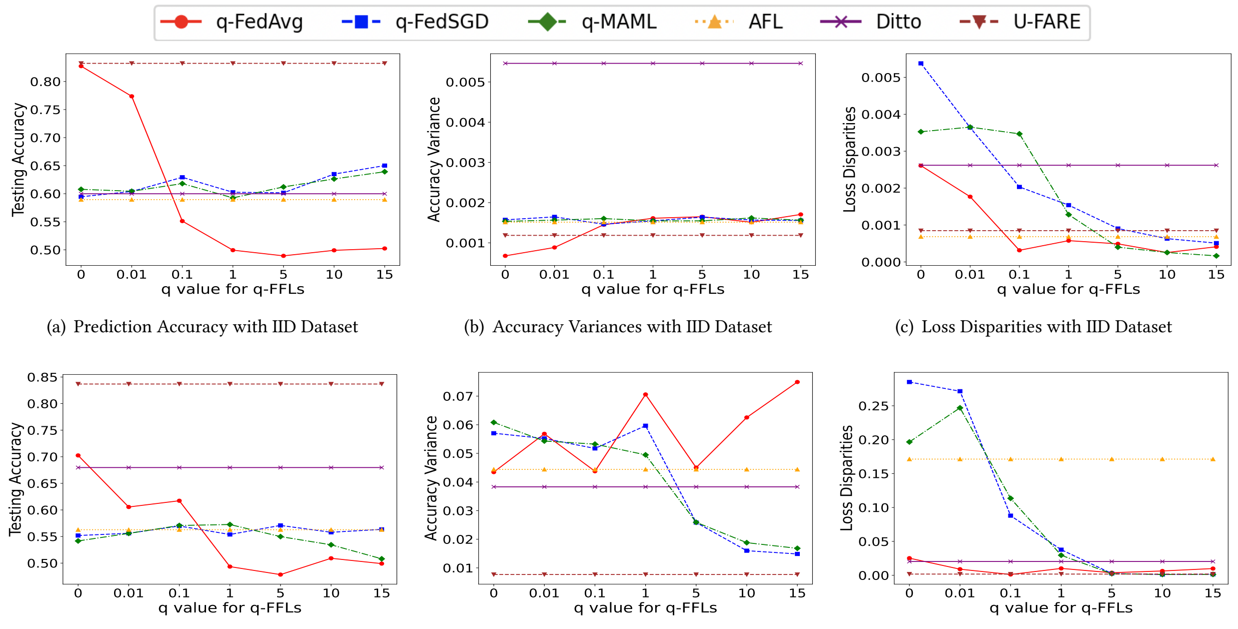
Figure 2: Effect of varying the level of fairness (q) under NACC IID and non-IID datasets.

model and dataset) to achieve the desired balance, where higher q enhances fairness.

7.2.2 **Varying the number of clients.** Figures 5 to 7 explore how performance metrics change with varying numbers of clients in the FL system. We find that U-FARE continues to outperform baseline methods across all three metrics, with no significant sensitivity to increasing the number of clients. Notably, for the ADNI dataset shown as Figure 7, Ditto demonstrates a more pronounced sensitivity to changes in client numbers across all metrics. As the number of clients increases, Ditto improves prediction accuracy and fairness, surpassing U-FARE when the client count exceeds 15. This suggests that Ditto's personalized approach benefits from larger-scale FL systems, leveraging the increased number of clients effectively.

7.2.3 **Varying the attack severity.** We investigate two types of adversarial attacks outlined in our threat model (see Section 4.3). We employ a backdoor approach for the poisoning attack by injecting triggers into 10% of the training data, where the triggers correspond to the positive class. To evaluate the impact of the Byzantine attack, we examine the attack severity, which represents the probability of an attacker successfully executing an attack at any given time t. For example, with an attack severity of 0.1, there is a 10% chance that the attacker will successfully inject Gaussian noise into the local model parameters. We experiment with the NACC dataset by injecting backdoor triggers into the dataset and then generating both IID and non-IID data as training samples.

Figure 8 illustrates the impact of different levels of attack severity, where higher severity triggers an attack to succeed more often, representing a higher impact on the performance metrics of FL schemes. For U-FARE, increased attack severity leads to a decline in prediction accuracy and fairness, in terms of AV, as illustrated in Fig. 8 (a), (b), (d) and (e). An increase in attack severity results in lower prediction accuracy due to the inclusion of more compromised local updates, undermining the model's prediction accuracy and AV performance. However, attack severity does not significantly impact LD, showing the system's robustness under one fairness metric. It is worth mentioning that Ditto is most robust to adversarial attacks under the IID setting in terms of fairness performance, as shown in Figue 8. As attack severity increases, the AV decreases, resulting in a fairer FL system. This is because Ditto trains personalized models for each client by considering both local data and shared global knowledge. This allows each client to maintain its model tailored to its data, reducing the impact of malicious updates from other clients.



(a) Prediction Accuracy with IID Dataset

(b) Accuracy Variances with IID Dataset

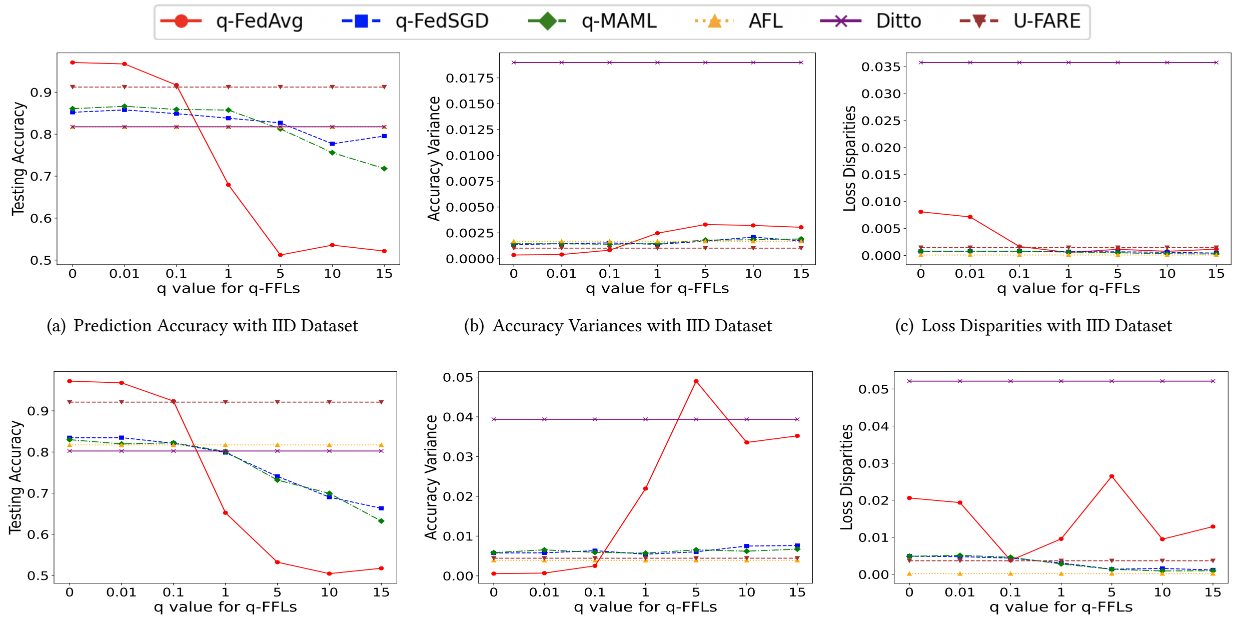
(c) Loss Disparities with IID Dataset

(d) Prediction Accuracy with Non-IID Dataset

(e) Accuracy Variances with Non-IID Dataset

(f) Loss Disparities with Non-IID Dataset

Figure 3: Effect of varying the level of fairness (q) under OASIS IID and non-IID datasets.



(a) Prediction Accuracy with IID Dataset

(b) Accuracy Variances with IID Dataset

(c) Loss Disparities with IID Dataset

(d) Prediction Accuracy with Non-IID Dataset

(e) Accuracy Variances with Non-IID Dataset

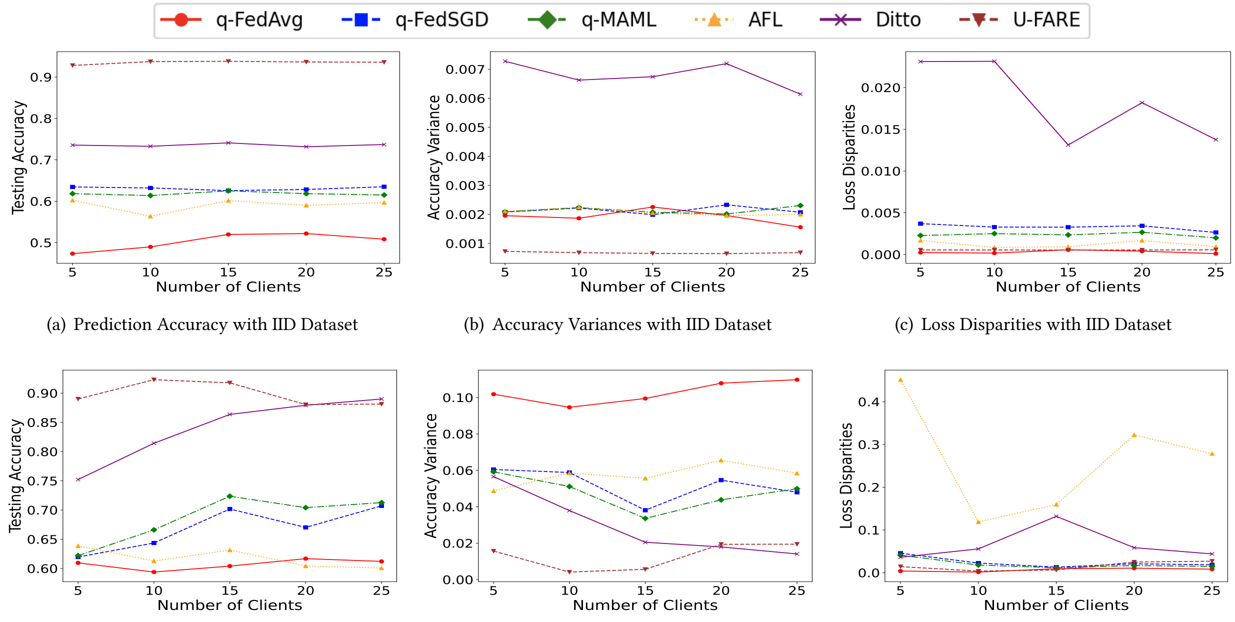
(f) Loss Disparities with Non-IID Dataset

Figure 4: Effect of varying the level of fairness (q) under ADNI IID and non-IID datasets.

8 CONCLUSION & FUTURE WORK

This work proposed U-FARE, an uncertainty-aware fair FL framework to enhance the prediction accuracy and fairness of AI models in healthcare systems, particularly in the context of Alzheimer's disease detection. U-FARE integrates evidential neural networks into the FL paradigm to quantify and manage uncertainty, offering significant advantages over existing fairness-aware FL methods.

The experimental results demonstrated the superior performance of U-FARE across a range of healthcare datasets, including NACC, OASIS, and ADNI, where it consistently outperformed other fairness-aware FL methods in terms of both prediction accuracy and fairness metrics. Our method maintained the lowest values for accuracy



(a) Prediction Accuracy with IID Dataset

(b) Accuracy Variances with IID Dataset

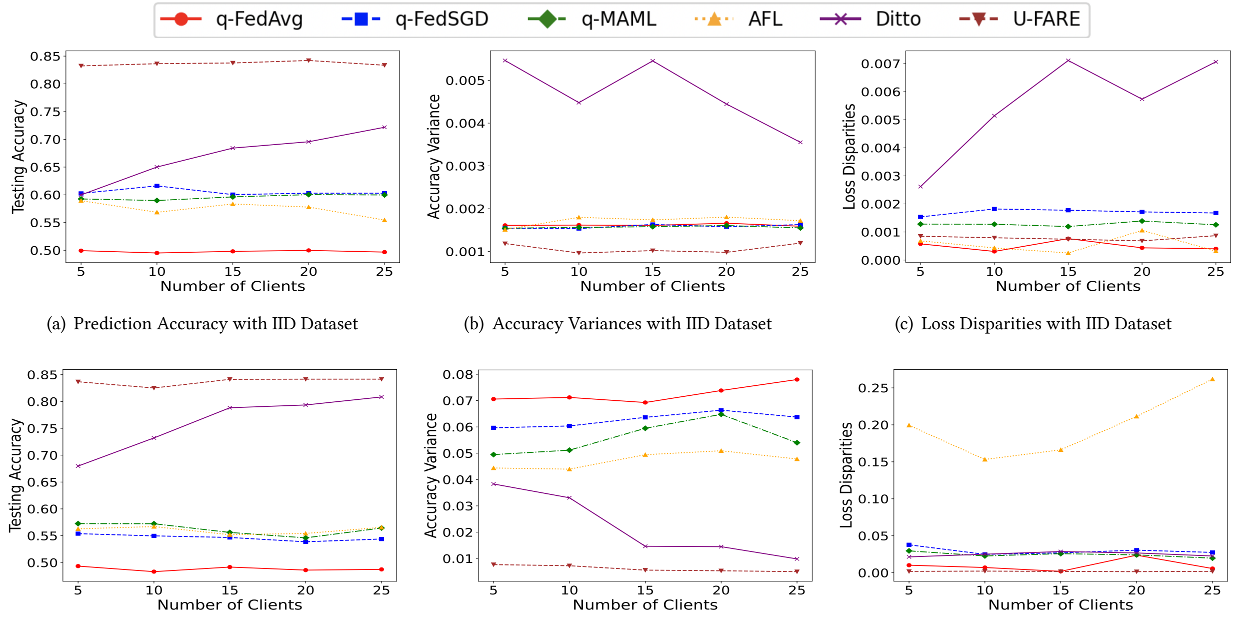
(c) Loss Disparities with IID Dataset

(d) Prediction Accuracy with Non-IID Dataset

(e) Accuracy Variances with Non-IID Dataset

(f) Loss Disparities with Non-IID Dataset

Figure 5: Effect of varying the number of clients under NACC IID and non-IID datasets.



(a) Prediction Accuracy with IID Dataset

(b) Accuracy Variances with IID Dataset

(c) Loss Disparities with IID Dataset

(d) Prediction Accuracy with Non-IID Dataset

(e) Accuracy Variances with Non-IID Dataset

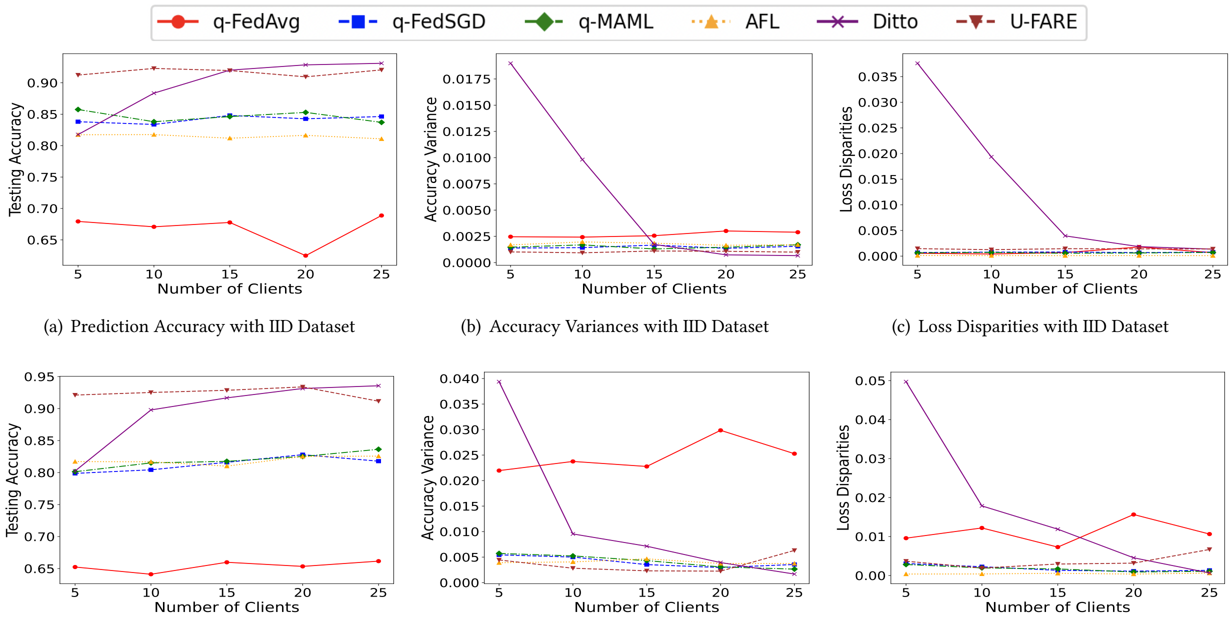
(f) Loss Disparities with Non-IID Dataset

Figure 6: Effect of varying the number of clients under OASIS IID and non-IID datasets.

variance and loss disparity, indicating its ability to reduce fairness issues and promote equitable performance across clients. Furthermore, U-FARE exhibited strong resilience to adversarial attacks, maintaining reliability in prediction accuracy and fairness even under poisoning and Byzantine attacks.

The insights gained from our sensitivity analysis suggest that while increasing fairness levels can reduce prediction accuracy, U-FARE balances these trade-offs effectively. Our approach, driven by uncertainty-aware principles, proves more reliable and robust than traditional fairness metrics like loss-based approaches.

For the **future work**, we aim to leverage privacy-preserving techniques into U-FARE to ensure both privacy and fairness while maintaining high prediction accuracy. We will also explore the relationship between fairness, privacy, and uncertainty to develop an



(a) Prediction Accuracy with IID Dataset

(b) Accuracy Variances with IID Dataset

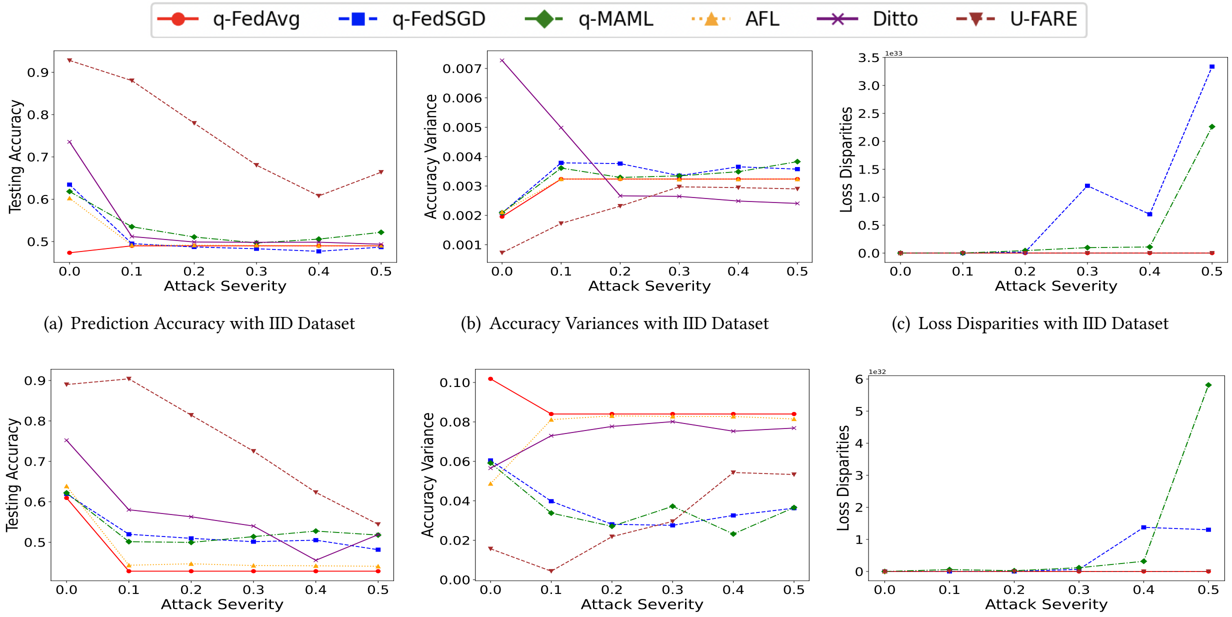
(c) Loss Disparities with IID Dataset

(d) Prediction Accuracy with Non-IID Dataset

(e) Accuracy Variances with Non-IID Dataset

(f) Loss Disparities with Non-IID Dataset

Figure 7: Effect of varying the number of clients under ADNI IID and non-IID datasets.



(a) Prediction Accuracy with IID Dataset

(b) Accuracy Variances with IID Dataset

(c) Loss Disparities with IID Dataset

(d) Prediction Accuracy with Non-IID Dataset

(e) Accuracy Variances with Non-IID Dataset

(f) Loss Disparities with Non-IID Dataset

Figure 8: Effect of varying the degree of attack severity under NACC IID and non-IID datasets.

optimal solution for addressing potential trade-offs. Furthermore, we plan to extend our work to handle multi-modal data, enhancing the performance and generalization of predictive models in healthcare scenarios.

ACKNOWLEDGMENT

This work is partly supported by NSF under grants 2107449, 2107450, and 2107451.

REFERENCES

[1] K Aditya Shastry and HA Sanjay. 2024. Artificial intelligence techniques for the effective diagnosis of Alzheimer's disease: a review. *Multimedia Tools and Applications* 83, 13 (2024), 40057–40092.

[2] Alzheimer's Disease Neuroimaging Initiative. 2024. ADNI | Alzheimer's Disease Neuroimaging Initiative. https://adni.loni.usc.edu/

[3] Eugene Bagdasaryan and Vitaly Shmatikov. 2021. Blind Backdoors in Deep Learning Models. In *30th USENIX Security Symposium (USENIX Security 21)*. USENIX Association, 1505–1521.

[4] Alex Beutel, Jilin Chen, Tulsee Doshi, Hai Qian, Allison Woodruff, Christine Luu, Pierre Kreitmann, Jonathan Bischof, and Ed H Chi. 2019. Putting fairness principles into practice: Challenges, metrics, and improvements. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*. 453–459.

[5] Peva Blanchard, El Mahdi El Mhamdi, Rachid Guerraoui, and Julien Stainer. 2017. Machine learning with adversaries: Byzantine tolerant gradient descent. *Advances in neural information processing systems* 30 (2017).

[6] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. 2002. SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research* 16 (2002), 321–357.

[7] Richard J Chen, Judy J Wang, Drew FK Williamson, Tiffany Y Chen, Jana Lipkova, Ming Y Lu, Sharifa Sahai, and Faisal Mahmood. 2023. Algorithmic fairness in artificial intelligence for medicine and healthcare. *Nature biomedical engineering* 7, 6 (2023), 719–742.

[8] Sen Cui, Weishen Pan, Jian Liang, Changshui Zhang, and Fei Wang. 2021. Addressing algorithmic disparity and performance inconsistency in federated learning. *Advances in Neural Information Processing Systems* 34 (2021), 26091–26102.

[9] Arthur P Dempster. 1968. A generalization of Bayesian inference. *Journal of the Royal Statistical Society: Series B (Methodological)* 30, 2 (1968), 205–232.

[10] Christoph Düsing and Philipp Cimiano. 2022. On the Trade-off Between Benefit and Contribution for Clients in Federated Learning in Healthcare. In *2022 21st IEEE International Conference on Machine Learning and Applications (ICMLA)*. IEEE, 1672–1678.

[11] Yahya H Ezzeldin, Shen Yan, Chaoyang He, Emilio Ferrara, and A Salman Avestimehr. 2023. Fairfed: Enabling group fairness in federated learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 37. 7494–7502.

[12] Biraja Ghoshal, Bhargab Ghoshal, and Allan Tucker. 2022. Leveraging uncertainty in deep learning for pancreatic adenocarcinoma grading. In *Annual Conference on Medical Image Understanding and Analysis*. Springer, 565–577.

[13] S Maryam Hosseini, Milad Sikaroudi, Morteza Babaie, and HR Tizhoosh. 2023. Proportionally fair hospital collaborations in federated learning of histopathology images. *IEEE transactions on medical imaging* (2023).

[14] Yibo Hu, Yuzhe Ou, Xujiang Zhao, Jin-Hee Cho, and Feng Chen. 2021. Multidimensional uncertainty-aware evidential neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35. 7815–7822.

[15] Audun Jøsang. [n. d.]. *Subjective logic*. Vol. 3. Springer.

[16] Audun Jøsang and Robin Hankin. 2012. Interpretation and fusion of hyper opinions in subjective logic. In *2012 15th International Conference on Information Fusion*. IEEE, 1225–1232.

[17] Kasem Khalil, Mohammad Mahbubur Rahman Khan Mamun, Ahmed Sherif, Mohamed Said Elsersy, Ahmad Abdel-Aliem Imam, Mohamed Mahmoud, and Maazen Alsabaan. 2023. A federated learning model based on hardware acceleration for the early detection of alzheimer's disease. *Sensors* 23, 19 (2023), 8272.

[18] Rajesh Kumar, Abdullah Aman Khan, Jay Kumar, Noorbakhsh Amiri Golilarz, Simin Zhang, Yang Ting, Chengyu Zheng, Wenyong Wang, et al. 2021. Blockchain-federated-learning and deep learning models for covid-19 detection using ct imaging. *IEEE Sensors Journal* 21, 14 (2021), 16301–16314.

[19] PJ LaMontagne, TL Benzinger, JC Morris, S Keefe, R Hornbeck, C Xiong, E Grant, J Hassenstab, K Moulder, AG Vlassenko, et al. 2019. OASIS-3: Longitudinal Neuroimaging, Clinical, and Cognitive Dataset for Normal Aging and Alzheimer Disease. (2019).

[20] Jiachun Li, Yan Meng, Lichuan Ma, Suguo Du, Haojin Zhu, Qingqi Pei, and Xuemin Shen. 2022. A Federated Learning Based Privacy-Preserving Smart Healthcare System. *IEEE Transactions on Industrial Informatics* 18, 3 (2022), 2021–2031. https://doi.org/10.1109/TII.2021.3098010

[21] Tian Li, Shengyuan Hu, Ahmad Beirami, and Virginia Smith. 2021. Ditto: Fair and robust federated learning through personalization. In *International conference on machine learning*. PMLR, 6357–6368.

[22] Tian Li, Maziar Sanjabi, Ahmad Beirami, and Virginia Smith. 2019. Fair Resource Allocation in Federated Learning. In *International Conference on Learning Representations*.

[23] Xiaoli Li, Siran Zhao, Chuan Chen, and Zibin Zheng. 2023. Heterogeneity-aware fair federated learning. *Information Sciences* 619 (2023), 968–986.

[24] Zelei Liu, Yuanyuan Chen, Yansong Zhao, Han Yu, Yang Liu, Renyi Bao, Jinpeng Jiang, Zaiqing Nie, Qian Xu, and Qiang Yang. 2022. Contribution-aware federated learning for smart healthcare. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 36. 12396–12404.

[25] Samual MacDonald, Kaiah Steven, and Maciej Trzaskowski. 2022. Interpretable AI in healthcare: Enhancing fairness, safety, and trust. In *Artificial Intelligence in Medicine: Applications, Limitations and Future Directions*. Springer, 241–258.

[26] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. 2017. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*. PMLR, 1273–1282.

[27] Syed Irfan Ali Meerza, Zhuohang Li, Luyang Liu, Jiaxin Zhang, and Jian Liu. 2022. Fair and Privacy-Preserving Alzheimer's Disease Diagnosis Based on Spontaneous Speech Analysis via Federated Learning. In *2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*. IEEE, 1362–1365.

[28] Angela Mitrovska, Pooyan Safari, Kerstin Ritter, Behnam Shariati, and Johannes Karl Fischer. 2024. Secure federated learning for Alzheimer's disease detection. *Frontiers in Aging Neuroscience* 16 (2024), 1324032.

[29] Mehryar Mohri, Gary Sivek, and Ananda Theertha Suresh. 2019. Agnostic federated learning. In *International conference on machine learning*. PMLR, 4615–4625.

[30] Kanchan Naithani, YP Raiwani, Shrikant Tiwari, and Alok Singh Chauhan. 2024. Artificial Intelligence Techniques Based on Federated Learning in Smart Healthcare. In *Federated Learning for Smart Communication using IoT Application*. Chapman and Hall/CRC, 81–108.

[31] Eric W Prince, Debashis Ghosh, Carsten Görg, and Todd C Hankinson. 2023. Uncertainty-aware deep learning classification of adamantinomatous craniopharyngioma from preoperative mri. *Diagnostics* 13, 6 (2023), 1132.

[32] Zixuan Qin, Liu Yang, Fei Gao, Qinghua Hu, and Chenyang Shen. 2022. Uncertainty-aware aggregation for federated open set domain adaptation. *IEEE Transactions on Neural Networks and Learning Systems* (2022).

[33] Joseph Rance and Filip Svoboda. 2023. Attacks of fairness in Federated Learning. *arXiv preprint arXiv:2311.12715* (2023).

[34] Shafiq Ul Rehman, Noha Tarek, Caroline Magdy, Mohammed Kamel, Mohammed Abdelhalim, Alaa Melek, Lamees N Mahmoud, and Ibrahim Sadek. 2024. AI-based tool for early detection of Alzheimer's disease. *Heliyon* 10, 8 (2024).

[35] Murat Sensoy, Lance Kaplan, and Melih Kandemir. 2018. Evidential deep learning to quantify classification uncertainty. *Advances in neural information processing systems* 31 (2018).

[36] Silvia Seoni, Vicnesh Jahmunah, Massimo Salvi, Prabal Datta Barua, Filippo Molinari, and U Rajendra Acharya. 2023. Application of uncertainty quantification to artificial intelligence in healthcare: A review of last decade (2013–2023). *Computers in Biology and Medicine* (2023), 107441.

[37] Saurabh Singh, Shailendra Rathore, Osama Alfarraj, Amr Tolba, and Byungun Yoon. 2022. A framework for privacy-preservation of IoT healthcare data using Federated Learning and blockchain technology. *Future Generation Computer Systems* 129 (2022), 380–388.

[38] Pegah Tabarisaadi, Abbas Khosravi, and Saeid Nahavandi. 2022. Uncertainty-aware skin cancer detection: The element of doubt. *Computers in Biology and Medicine* 144 (2022), 105357.

[39] Yichen Wan, Youyang Qu, Wei Ni, Yong Xiang, Longxiang Gao, and Ekram Hossain. 2024. Data and model poisoning backdoor attacks on wireless federated learning, and the defense mechanisms: A comprehensive survey. *IEEE Communications Surveys & Tutorials* (2024).

[40] Meng Wang, Lianyu Wang, Xinxing Xu, Ke Zou, Yiming Qian, Rick Siow Mong Goh, Yong Liu, and Huazhu Fu. 2023. Federated Uncertainty-Aware Aggregation for Fundus Diabetic Retinopathy Staging. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 222–232.

[41] Abbas Yazdinejad, Ali Dehghantanha, and Gautam Srivastava. 2023. Ap2fl: auditable privacy-preserving federated learning framework for electronics in healthcare. *IEEE Transactions on Consumer Electronics* (2023).

[42] Li Zhang, Jianbo Xu, Pandi Vijayakumar, Pradip Kumar Sharma, and Uttam Ghosh. 2022. Homomorphic encryption-based privacy-preserving federated learning in iot-enabled healthcare system. *IEEE Transactions on Network Science and Engineering* (2022).