information

MDPI

Article

Trustworthy Federated Learning with Blockchain-Based Consensus for Mitigating Poisoning Attacks in Healthcare Systems

Raghad Hamed Alhamrani *, Fatmah Omar Bamashmoos and Enas Fawzi Khairallah

Department of Information Technology, Faculty of Computing & Information Technology, King Abdulaziz University, Jeddah 21589, Saudi Arabia; fbamashmoos@kau.edu.sa (F.O.B.); ekhairallah@kau.edu.sa (E.F.K.)
* Correspondence: ralhamrani0006@stu.kau.edu.sa

Abstract

This paper presents a framework that integrates blockchain-enabled Federated Learning (FL) with consensus mechanisms to mitigate poisoning attacks in healthcare environments. The framework incorporates blockchain consensus mechanisms, with Proof-of-Work (PoW) used as a baseline and Proof-of-Stake (PoS) adopted as the proposed approach; both are evaluated independently within the same Secure Multiparty Computation (SMPC)-enabled federated learning architecture for privacy preservation. The proposed system is evaluated on the OCTMNIST and TissueMNIST datasets under both centralized and federated settings, including poisoning scenarios with 10% and 50% malicious clients. Results show that consensus-aware aggregation reduces the influence of unreliable client updates and improves the robustness of the global model under poisoning conditions. In addition, the framework prioritizes trustworthy client contributions during aggregation, supporting reliable model sharing in collaborative healthcare learning environments. Unlike prior blockchain-based federated learning defenses that introduce heavy cryptographic overhead, the proposed PoS-based aggregation explicitly balances robustness and computational efficiency, enabling practical deployment under high poisoning ratios.

Keywords: federated learning; blockchain; information security; trust management; secure aggregation; healthcare data; poisoning attacks

1. Introduction

Artificial Intelligence (AI) has been increasingly adopted across various sectors, with healthcare emerging as one of the fastest-growing domains. Recent AI advancements have demonstrated significant potential to improve diagnostic accuracy, clinical decision-making, and overall healthcare efficiency. However, the adoption of AI in healthcare introduces critical challenges, particularly related to data privacy, data security, and model robustness [1,2].

Federated Learning (FL) is an evolving approach in decentralized Machine Learning (ML), enabling simultaneous model training across numerous devices or clients without necessitating data sharing. FL has gained recognition as an effective solution for mitigating data privacy concerns [3].

Ensuring the integrity and trustworthiness of shared information is a fundamental challenge in distributed learning systems, particularly in collaborative healthcare environments where unreliable updates can directly degrade decision quality and patient safety.

Check for updates

Academic Editor: Aristeidis Karras

Received: 14 January 2026
Revised: 4 February 2026
Accepted: 10 February 2026
Published: 14 February 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

Although FL improves data privacy by keeping data local, it does not inherently prevent poisoning attacks, where participants manipulate local data or model updates to degrade the global model's performance. While blockchain can reduce risks associated with a centralized coordinator, it does not guarantee that client-submitted updates are trustworthy; therefore, blockchain-based FL remains vulnerable to client-side attacks [4,5].

FL produces and aggregates a global model from multiple participants or clients. The poisoning attacks can compromise the accuracy and reliability of healthcare models, leading to misdiagnosis in clinical and scientific settings [6]. To address poisoning attacks, it is crucial to integrate additional techniques into FL training to filter out poisoned models and ensure global model aggregation against malicious manipulation.

This paper presents an advanced AI framework that integrates FL, blockchain technology, and consensus mechanisms to prevent malicious attacks in healthcare scenarios. The integration of blockchain and consensus technologies can enhance data privacy preservation, support performance-based client participation control, and improve robustness against poisoning attacks, thereby enhancing the performance and accuracy of AI systems. By utilizing datasets from the MedMNIST collection (medmnist Python package, version 3.0.2), including OCTMNIST and TissueMNIST , the experiments were conducted to simulate real-world healthcare scenarios [7].

Blockchain's inherent features, such as transparency, immutability, and decentralized control, enhance the security and reliability of FL models [8]. Furthermore, consensus mechanisms such as Proof-of-Work (PoW) and Proof-of-Stake (PoS) provide an effective solution for securing model aggregation by supporting secure, verifiable aggregation while enhancing integrity and trust in the federated learning process [9,10].

The methodology in this study used an FL approach and compared it with centralized learning across three scenarios. In the first scenario, data was divided among ten clients, and the global model was updated through several rounds. In the second scenario, 10% and 50% poisoning attacks were simulated to see how they would affect the global model's performance. A Secure Multiparty Computing (SMPC) technique was employed to preserve the confidentiality of client model updates during the aggregation process [11]. The third scenario integrates FL with blockchain technology and consensus mechanisms to ensure robust and reliable model aggregation among participating clients.

In the proposed framework, the aggregated global model is recorded on the blockchain using a PoS consensus mechanism [12,13]. PoS enables performance-aware participation by allowing clients with higher accuracy-based stakes to have greater influence during aggregation. This performance-aware aggregation process suppresses unstable or low-quality updates, thereby reducing the influence of malicious or unreliable clients on the global model.

Despite recent advances in blockchain-enabled federated learning, most existing solutions primarily emphasize security guarantees without adequately addressing the trade-off between robustness and computational efficiency, particularly at high poisoning ratios. In healthcare environments, where scalability and timely model updates are critical, excessive cryptographic overhead can limit practical deployment. To address this gap, this work introduces a performance-weighted PoS aggregation strategy, while SMPC is employed to protect local model updates, enabling robust mitigation of poisoning attacks while significantly reducing training and aggregation time. The proposed framework is evaluated under severe attack scenarios (up to 50% malicious clients) using real medical datasets, highlighting its practical applicability.

Unlike existing blockchain-based federated learning approaches that primarily emphasize security guarantees, this work explicitly investigates the trade-off between robustness and computational efficiency under high poisoning ratios. By incorporating performance-

weighted Proof-of-Stake directly into the aggregation process rather than treating PoS as an auxiliary incentive or reputation layer, and by reporting a detailed time-complexity analysis, the proposed framework demonstrates that robustness against adversarial clients can be achieved without incurring excessive computational overhead, which is critical for practical healthcare deployments.

The main contributions of this work are summarized as follows:

1. Design of a blockchain-enabled federated learning architecture tailored for healthcare systems, integrating SMPC-based privacy preservation with consensus-based global aggregation.
2. Proposal of a performance-weighted Proof-of-Stake (PoS) mechanism that prioritizes reliable client updates based on historical accuracy.
3. Comprehensive evaluation of robustness against poisoning attacks under both moderate (10%) and severe (50%) malicious client scenarios using medical imaging datasets.
4. Comparative analysis of accuracy, robustness, and computational efficiency, highlighting the practical feasibility of PoS-based aggregation compared to PoW baseline under high poisoning ratios.

The remainder of this paper is structured as follows: Section 2 analyzes relevant previous studies. Section 3 discusses the proposed methodology. Section 4 presents and discusses the findings. Finally, Section 5 provides a conclusion.

2. Related Work

This section discusses the need to improve the quality of medical environments, including data privacy, time-to-processing, and client training in AI. For instance, Mohammed et al. [14] introduced a model called EDFOS to enhance data privacy and blockchain-based networks through simulations. The EDFOS system reduced power consumption by 39% compared to existing healthcare systems and cut training and testing time by 29%. In addition, as the ratio of malicious clients increases, the performance of FL models significantly drops.

Kalapaaking et al. [15] proposed a blockchain-based FL model that incorporates SMPC to enhance against poisoning attacks. Using the OCTMNIST and TissueMNIST datasets with a ResNet-18 CNN model enables secure model verification via SMPC, eliminating compromised local models. Once local models are verified, they are sent to blockchain nodes for secure aggregation, and the global model is stored in a tamper-proof storage system. The average poisoned-client accuracy declined to 30% on the OCTMNIST datasets when 10% of clients were malicious, while the global evaluation accuracy dropped by 7%. Global model accuracy could decline 22% with 50% malicious clients. 10% malicious clients reduced global model accuracy by 9% for TissueMNIST, whereas 50% reduced it by 26%.

N. Dong et al. [16] suggested a way to protect FL from poisoning attacks by combining blockchain technology with a stake-based aggregation mechanism. The suggested FedAVG with blockchain worked just as well as FedAVG without malicious attacks, and it worked better than FedAVG with malicious attacks all the time when using the ChestX-ray14 dataset. The results showed that the proposed stake-based aggregation mechanism effectively identifies and stops bad behavior in FL settings.

C. Gan et al. [17] proposed a dual-blockchain strategy to eliminate low-quality nodes and reduce poisoning attacks. It used the MedMNIST dataset on two blockchains, MQchain and RIchain. The proof-of-quality consensus algorithm in MQchain keeps out low-quality nodes, while the reputation evaluation mechanism in RIchain does the same. Compared to three other methods, the model made it harder for attacks to work.

In addition, another study [18] proposed an approach named PBSL to resist poisoning attacks while safeguarding data privacy with incorporated threshold fully homomorphic

encryption and the MNIST and FashionMNIST datasets. It was developed by integrating deep learning techniques with blockchain-based smart contract platforms. PBSl showed consistent accuracy, with only a slight drop from 98.0% at 10% malicious MCs to 96.0% at 50%. The limitation of this study is the diagnostic accuracy challenges caused by insufficient data sharing.

L. Tian et al. [19] proposed an AI model named PEFL, which leverages blockchain to facilitate privacy protection and coordination among customers. PEFL incorporated an aggregation-aspect detection set of rules to counter poisoning attacks and proposed the MFF consensus mechanism, which was evaluated on the MNIST and CIFAR-10 datasets. Their model executed better training efficiency while ensuring privacy and security. The proposed framework nonetheless confronted challenges in real-world applications, despite its upgrades.

C. Wan et al. proposed a decentralized FL that combined blockchain technology with distillation protection [20]. A rotation mechanism randomly assigned clients to roles so that awful clients could not be assigned to the same roles. This feature makes it easier for the local purchaser to counter attacks with adversarial samples, reduces communication overhead, accelerates training, and improves the model's generalization. Even as the number of adversarial samples increased, the proposed model trained with an accuracy of over 0.880. The limitation emphasized is that it functions solely within a simulated environment.

Bharath et al. [21] proposed a blockchain-based federated learning model named ATB-FL. This recent preprint integrates behavior-based trust evaluation with blockchain authentication and incentive–penalty mechanisms, and was evaluated on the CIC-IoMT 2024 dataset.The ATB-FL framework reported a diagnostic accuracy of 95.1% and reduced the misclassification rate to below 5%. However, the evaluation was limited to a single dataset, raising concerns regarding scalability and generalizability.

M. Xu and X. Li [22] employed a model named FedG2L to address poisoning attacks and single-point failures in centralized FL using the Gradient-Similarity-Based Secure Consensus Algorithm. The model is based on gradient similarity, enabling the identification and removal of gradients that deviate, using an improved auxiliary classifier generative adversarial network for data generation. Accuracy improved by at least 55% compared with schemes that did not incorporate any defense mechanisms. The attack success rate was reduced by more than 60% in mitigating poisoning attacks. The strategy assumed that some groups were equally limited rather than completely malicious and possessed unlimited power. This restricts how strong the attacker model may be.

W. Liu, et al. [23] developed the BFG model based on blockchain and FL using the interplanetary file system, differential privacy, and generative adversarial networks with a consensus protocol of PoS. BFG reduces the success rate of poisoning attacks. For 10% poisoning, attackers achieved approximately a 26% attack success rate on the MNIST and 23% on the CIFAR-10 dataset. The limitations were due to the use of partial and limited images.

R. Myrzashova et al. [24] developed a framework based on blockchain and FL to identify 15 distinct lung diseases and the NIH Chest X-Ray dataset. The model evaluated on a test accuracy of 92.86% and a latency of 43.518 ms. It demonstrated robustness, with resilience metrics consistently approaching 87% across three evaluated cyberattacks.

G. Sun et al. [25] proposed an attack on federated learning (AT2FL) to address high communication costs and stragglers using computing gradients for the target nodes. Endoscopic images, human activity recognition, and the Parkinson's dataset were used in this study. The EndAD and Human Activity datasets showed performance deterioration of 21.707% and 26.836% in classifying direct attacks, respectively.

Z. Ma et al. [26] proposed SFPA (Secure Federated Learning against Poisoning Attacks), a secure federated learning framework designed to provide privacy-preserving random

forest-based FL across multiple data islands and a multi-key decryption scheme. SFPA utilizes RF as the ensemble learner for medical diagnosis. RF models are chosen for their interpretability and transparency, consisting of a series of decision trees. Experiments showed that the accuracy of SFPA with secure defense ranged from 84.3–97.4% even with increased poisoning. However, the key generation center assumption could be a point of vulnerability if the KGC were compromised.

Previous studies have employed various methodologies to address poisoning attacks in FL. Some studies focused on FL without addressing filtering poisoning attacks, whereas others proposed filtering techniques to enhance the security of the global model, as demonstrated in Table 1. A comprehensive comparison of federated learning (FL) and attack-related studies in healthcare is provided in Supplementary Table S1. However, filtering methods usually increase training time, especially in the cloud. Small datasets limited some studies, making it challenging to generalize their results to larger contexts.

Table 1. Blockchain-based federated learning defense and attack-resilient methods in healthcare.

Reference	Method	Key Results	Remarks
Mohammed et al. (2023) [14]	EDFOS (Blockchain FL)	39% power reduction; 29% faster training	Energy efficient; simulation only
Dong et al. (2024) [16]	Stake-based blockchain FL	Outperforms FedAvg under poisoning	Single-dataset validation
Gan et al. (2025) [17]	Dual-blockchain FL	Improved poisoning resilience	High system complexity
Zhu et al. (2024) [18]	PBSL (Blockchain + FHE)	Accuracy drop: 98% to 96% (50% attackers)	Strong privacy; scalability issues
Tian et al. (2025) [19]	PEFL	Improved efficiency and security	Deployment challenges
Kalapaaking et al. (2024) [15]	Blockchain FL + SMPC	30% acc. loss (10% attackers); 50% (50%)	Secure but accuracy-sensitive
Wan et al. (2024) [20]	Blockchain + distillation	Higher accuracy; adversarial robustness	Simulation only
B. M. B. et al. (2025) [21]	ATB-FL	95.1% diagnostic accuracy	Scalability limits
Xu and Li (2023) [22]	FedG2L	55% acc. gain; 60% attack reduction	Limited threat model

In this study, we aim to contribute by utilizing FL and assessing various learning strategies, including centralized and federated settings under label-flipping attacks at 10% and 50%. The framework incorporates SMPC-based privacy preservation and blockchain consensus mechanisms to improve robustness and efficiency. We expect improved performance and reduced training time, addressing limitations in prior work and supporting practical healthcare applications.

3. Methodology

This paper proposes a FL architecture based on blockchain technology and a consensus mechanism, and evaluates it alongside SMPC-based secure aggregation for privacy preservation. This methodology aims to develop healthcare environments in which shared machine learning models are trained for multiple clients or hospitals while preserving data privacy and model integrity across different medical facilities. Figure 1 illustrates the proposed methodology, in which each hospital's local model is trained independently on its own data without sharing it with other clients. Secure verification is then performed between hospital clients and the cloud. SMPC-based secure aggregation is then performed to preserve the confidentiality of client updates, while poisoning attacks are mitigated during the aggregation process. Blockchain technology based on consensus mechanisms is applied to provide tamper-resistant storage and coordinated model dissemination, supporting the evaluation of system efficiency and robustness.

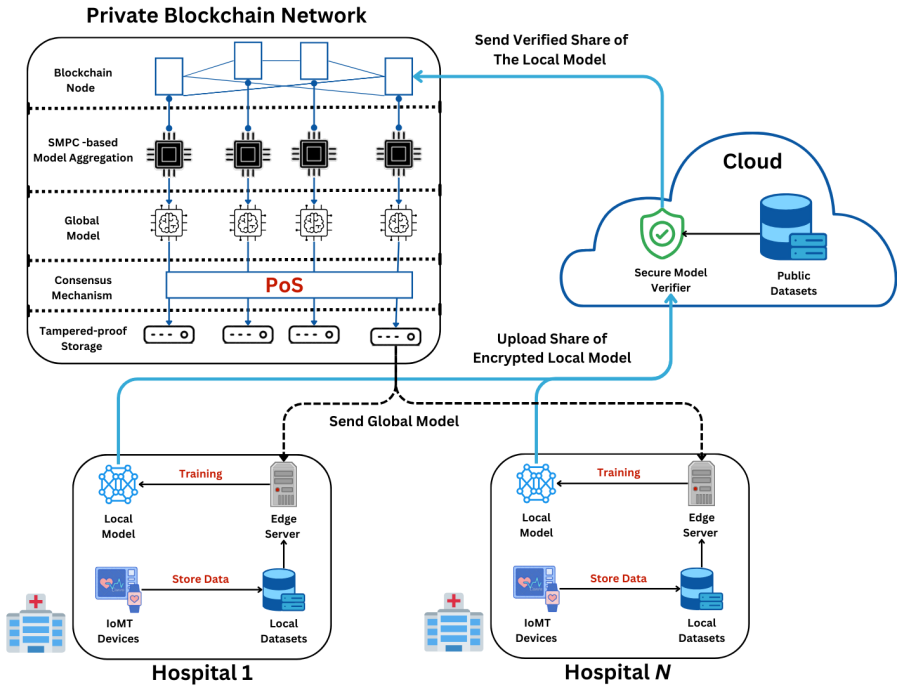


Figure 1. The proposed PoS-based blockchain federated learning workflow.

3.1. Dataset Description

The present research utilizes benchmark datasets from the MedMNIST collection [7], a compilation of preprocessed biomedical image datasets frequently used to evaluate machine learning models in medical classification tasks. The MedMNIST datasets are designed for rapid and efficient evaluation. They include standardized 28 × 28 grayscale images accompanied by classification labels from various domains and degrees of complexity.

From this set, two datasets, TissueMNIST and OCTMNIST, were selected for training and evaluating agents across the proposed FL environment. Both datasets were chosen for their suitable sample sizes for AI and their clinical relevance. The datasets are also multiclass and suitable for simulating real-world, data-intensive healthcare scenarios in a federated environment.

TissueMNIST contains 236,386 human kidney cortical cell samples collected from three reference tissue samples. The dataset contains eight diagnostic classes. OCTMNIST contains 109,309 optical coherence tomography images used by physicians to identify eye diseases. It is also considered a multiclass classification because it has four diagnostic classes.

3.2. Preprocessing and Splitting

In all experiments, both datasets are evenly divided among ten simulated clients, each representing a different hospital in the FL system. This configuration emulates a cross-silo healthcare federated learning setting (e.g., multiple hospitals or clinical sites) while maintaining computational feasibility under blockchain and consensus overhead. Moreover, selecting ten clients enables controlled poisoning scenarios, where 10% corresponds to one malicious client and 50% to five malicious clients, ensuring consistent comparisons across aggregation and consensus mechanisms.

Each client trains a local model on its own data partition without sharing raw data, reflecting practical privacy constraints in cross-institutional healthcare collaborations. Table 2 summarizes the dataset characteristics and the overall data split used in the experiments.

Table 2. Dataset sampling splitting for training, validation, and testing.

Dataset	Total Samples	Training Set	Validation Set	Testing Set
TissueMNIST	236,386	165,466	23,640	47,280
OCTMNIST	109,309	97,477	10,832	1000

3.3. Centralized Learning

The ResNet-18 model is used as the main CNN model in this study [27]. ResNet-18 has 18 layers and about 11 million parameters, offering a deep architecture that has been shown to be effective across various image classification tasks. This model can also load a pretrained version, which has been trained on a large dataset such as ImageNet, to potentially improve performance on the target task. Given that ResNet-18 expects RGB input images with three color channels, we convert grayscale images into RGB format by replicating the grayscale channel across the three RGB channels. The training process is carried out with a batch size of 64, Adam optimizer [28] with an initial learning rate of 0.001, and 40 epochs, followed by updates to the model's weights.

3.4. FL Architecture

In FL scenarios, several clients collaborate to build a global model while maintaining data segregation [28]. The federated training process is conducted over multiple communication rounds. In each round, clients independently train local models on their respective datasets and share only the model weights to enhance the global model [15]. The initial phase involves establishing a global model based on the ResNet-18 architecture. The global model weights are first synchronized among all client models. This ensures that each client commences with an identical initial state. The CrossEntropyLoss function is used to compute the loss during local training, and the Adam optimizer is used to update the model weights. The model weights that have been trained are sent back to the central server, as shown in Table 3.

Table 3. The hyperparameter of the proposed FL approaches.

Variables	Description
Number of Clients	10
Data Splitting per Client	80%, 10% and 10%
Number of Labels	4 in OCTMNIST, and 8 in TissueMINIST
Learning Rate	0.001
Batch Size	32
Epochs	10
Rounds	40

Upon completion of local training for each client, the client model is evaluated on a common test set, thereby establishing a uniform performance baseline across all clients. Subsequently, the server consolidates the weights of all client models by Federated Averaging (FedAvg). The server computes the mean of the received model weights and updates the global model accordingly. The updated global model is then evaluated on the same test set to assess overall performance across clients.

3.5. Threat Model and Assumptions

In a realistic healthcare setting, multiple hospitals collaboratively train diagnostic models while retaining local patient data (cross-silo federated learning across clinical institutions). In such environments, compromised participants can degrade global models

(e.g., by submitting unreliable or poisoned local updates), motivating the need for robust aggregation mechanisms, such as those proposed in this work.

This work considers a semi-honest but potentially malicious federated learning environment. The attacker is assumed to control a subset of participating clients, set to 10% and 50% in the experimental scenarios, and can perform label-flipping poisoning attacks by manipulating local training data. Malicious clients may attempt to degrade the global model by submitting corrupted updates; however, they do not have control over the blockchain layer or the consensus mechanism.

Collusion among malicious clients is limited to the defined percentage, and attackers cannot access or manipulate encrypted model updates generated through SMPC. The blockchain network is assumed to be permissioned and correctly implemented, ensuring that consensus protocols execute as intended.

This work assumes a permissioned blockchain setting in which consensus nodes are not fully compromised. Attacks targeting the blockchain layer itself, such as a majority takeover or validator collusion, are considered out of scope. Additionally, while stronger poisoning strategies such as model replacement and fully adaptive Byzantine attacks exist, this study focuses on label-flipping attacks as a representative and prevalent threat in healthcare federated learning systems. Addressing more advanced adversarial behaviors is left for future work.

3.6. Malicious Clients and SMPC-Based Secure Aggregation

In an FL environment, clients may diverge from the established protocol, which could trigger various attacks that jeopardize model integrity and data privacy [29].

Certain clients may intentionally alter labels in their local training data to degrade the model's overall performance. Others may employ sign or scaling manipulations to alter their gradient updates, thereby disrupting the aggregation process and introducing bias into the global model. Sophisticated adversarial strategies encompass model substitution and backdoor implantation. In model replacement, a client substitutes its local model with one intended for malicious intentions, thereby diminishing the reliability of the global model [30].

In this work, we simulate the malicious clients and label-flipping attack twice: once with 10% of the clients malicious and again with 50% malicious, and then run the SMPC protocol for secure aggregation.

SMPC allows multiple parties to collaboratively compute a function based on their private inputs while keeping them confidential [31]. In the context of FL, SMPC protocols are used to securely aggregate client updates, ensuring that neither the server nor other clients learn any individual client's update. A commonly employed method for such encryption is additive secret sharing, where:

Each client C_i splits its update Δw_i into n shares $(\Delta w_i^{(1)}, \Delta w_i^{(n)})$ such that:

$$\Delta \omega_i = \sum_{j=0}^{n} \Delta \omega_i^{(j)} \pmod{q}, \tag{1}$$

The shares are then distributed to either the aggregator or peer clients, depending on the SMPC protocol topology. The global model update is computed as

$$\Delta \omega = \sum_{i \in S}^{n} \Delta \omega_i \tag{2}$$

where $S \subseteq C$ represents the subset of clients selected in a given round. Although SMPC ensures input confidentiality, it does not guarantee the correctness or honesty of client-submitted inputs.

In this work, SMPC principles are adopted for secure aggregation, while TenSEAL is utilized as a practical fully homomorphic encryption (FHE) tool to enable encrypted computation during the aggregation process [32]. Although TenSEAL relies on homomorphic encryption, it is employed here as a practical instantiation of secure multiparty computation for encrypted federated aggregation. TenSEAL enables encryption and decryption of model weights during the aggregation process, ensuring that the server and other clients do not gain access to individual client updates. This protocol enhances the security of FL, especially in scenarios where malicious clients may attempt to disrupt the model training process.

In the SMPC implementation for federated training, the TenSEAL context is initialized using the CKKS scheme with specific encryption parameters for secure computation. Each client trains a local model and computes the updates $\Delta\omega_{(i)}$, and encrypts them using TenSEAL, splitting the updates into smaller parts stored as TenSEAL tensors. The server aggregates the encrypted updates using the Federated Averaging (FedAvg) method, ensuring the updates remain secure. Finally, the server decrypts the aggregated weights and updates the global model, maintaining security without exposing any client-specific data.

3.7. Blockchain Consensus Mechanism

A consensus mechanism is a fault-tolerant system used in the blockchain to get all of the distributed nodes to agree on a single state of the network [33]. These mechanisms ensure that distributed nodes remain synchronized and agree on the validity of transactions recorded on the blockchain.

The blockchain and consensus mechanisms are implemented in a simulated permissioned environment to evaluate relative aggregation behavior and robustness rather than real-world cryptocurrency mining or deployment-level energy consumption.

3.7.1. Proof of Work (PoW)

Proof-of-work underpinned early blockchain networks. Verified and first-mined blocks are appended to the current block after the final block in the chain. The hash function, such as SHA-256, must output a target range to accept the block when the nonce, the previous block's hash, and the new block's transactions are used as input [34]. The nonce can only be obtained by repeatedly testing different nonce values until the hash function produces an output within the required range. A participant broadcasts the block and transactions to other nodes after finding the nonce. If the new block is verified and mined first after the last block, it will join the chain. PoW competitors race to identify the correct nonce. Solution-searching is a weighted random coin-tossing mechanism in which a player with a greater hash rate (computational power) may be selected to append the next block to the chain. The probability p_i that participant i leads a network of N participants is

$$p_i = \frac{c_i}{\sum_{j=1}^{N} c_j} \tag{3}$$

3.7.2. Proof of Stake (PoS)

The inaugural Proof-of-Stake (PoS) network, originally introduced as PPCoin (later known as Peercoin) [35], was proposed to reduce the computational and energy demands associated with Proof-of-Work (PoW).

Participants with greater coinage, defined as the product of network tokens and their holding duration, are more likely to be selected. Each node in Peercoin specifically

resolves a PoW puzzle with an individual difficulty level, which can be diminished via the utilization of coinage. Contemporary PoS networks eradicate solution searching and no longer appoint block leaders based on computational prowess. Their option is determined by the stakes they own. The stake-based leader selection procedure markedly reduces a node's probability of being chosen as a leader based on its computational capacity, thereby reducing the energy consumption of PoS processes compared to PoW. Furthermore, PoW networks maintain block generation and transaction confirmation rates at very low, constant levels to ensure security, as numerous distinct blocks are proposed by miners. Conversely, because only a single block is produced in each round of Proof-of-Stake systems, block production and transaction confirmation speeds are typically significantly faster, leading to the recent popularity of PoS mechanisms.

In this work, each client maintains a list of validation accuracy values obtained in previous federated learning rounds. If the list is empty (i.e., the client has not yet participated in any round), the stake is initialized to 1. Otherwise, the stake is computed as the average accuracy across all completed rounds.

This design follows the PoS principle, where each client's stake (and consequently its influence during aggregation) is determined by historical performance. In the federated learning context, clients that consistently achieve higher accuracy contribute more strongly during the aggregation process, analogous to PoS protocols that prioritize participants with higher stakes.

The detailed pseudocode of the proposed federated learning framework and the PoS-based stake calculation mechanism is presented in Algorithms 1 and 2. These algorithms formally describe the training, evaluation, and aggregation procedures adopted in this study.

Algorithm 1 Federated Learning with Blockchain and Consensus Mechanism

Input: Number of clients n, model parameter length M_{len}
Output: Global model GM
Initialization: Initialize global model GM, number of training rounds num_rounds, local weights list $local_weights$, local accuracy list $local_accuracies$, and client stakes $stakes$
for each round $r = 1$ to num_rounds **do**
  Reset $local_weights$ and $local_accuracies$
  for each client P_i, $i = 1$ to n **do**
    Initialize client model M_i using ResNet18
    Train local model M_i on client data
    Compute local accuracy acc_i
    Store acc_i in $local_accuracies[P_i]$
    Update client stake:

$$stakes[P_i] \leftarrow calculate_stake(acc_i)$$

    if $stakes[P_i] > 0.5$ **then**
      Add model weights of P_i to $local_weights$
      Add acc_i to aggregation accuracy list
    end if
  end for
  Aggregate global weights:

$$global_weights \leftarrow average_weights(local_weights, local_accuracies)$$

  Update global model:

$$GM \leftarrow load_state_dict(global_weights)$$

end for
Evaluate final global model on test data:

$$GM \leftarrow evaluate_model(GM, test_loader)$$

Threshold Sensitivity Analysis: The stake threshold $\pi > 0.5$ was selected to balance robustness and client inclusion. To provide an empirical justification for this choice, a sensitivity analysis was conducted by evaluating threshold values of 0.4, 0.5, and 0.6 under identical poisoning settings. As shown in Table 4, thresholds below 0.5 admitted unreliable updates and reduced accuracy, while higher thresholds limited client participation

without additional performance gains. Based on this analysis, the threshold value of 0.5 demonstrated a stable trade-off between robustness and client inclusion and was therefore adopted in this study.

Table 4. Sensitivity analysis of stake threshold values.

Stake Threshold (p_i)	Accuracy (%)
0.4	89.6 ± 1.2
0.5	93.4 ± 0.8
0.6	92.9 ± 1.0

Algorithm 2 Blockchain-based PoS (calculate_stake)

Input: Client accuracies list *client_accuracies*
Output: Calculated stake value *Stake*
Begin
if *Client_accuracies* is empty **then**
  $Stake \leftarrow 1$ ▷ Initial stake when no previous accuracies exist
  Return *Stake*
else
  $Stake \leftarrow \frac{\sum(client_accuracies)}{\text{length}(client_accuracies)}$
  Return *Stake*
end if
End

In Algorithms 1 and 2, n denotes the total number of participating clients, and P_i represents the i-th client, where $i = 1, 2, \ldots, n$. The variable r denotes the federated learning round index, while *num_rounds* represents the total number of training rounds. The parameter M_{len} refers to the length of the model parameters. The local model trained by client P_i is denoted by M_i, and GM denotes the global model.

The local validation accuracy obtained by client P_i after local training is denoted by acc_i. The list *local_weights* stores the model weights submitted by clients selected for aggregation in a given round. The container *local_accuracies* records accuracy values associated with client models and is used to derive aggregation-related decisions.

The variable $stakes[P_i]$ denotes the reputation-based stake assigned to client P_i, computed using the function *calculate_stake()*. This stake value reflects the historical performance of the client and is used as a criterion to determine its eligibility to participate in the aggregation process. The function *average_weights()* performs weighted aggregation of local model updates. The functions *load_state_dict()* and *evaluate_model()* are used to update and evaluate the global model GM, respectively, using the test dataset provided by *test_loader*.

The symbols and variables used in Algorithms 1 and 2 are summarized in Table 5.

Table 5. Performance results of centralized learning for both datasets. Reported values correspond to average performance obtained across repeated experimental runs.

Symbol	Description
n	Number of participating clients
P_i	The i-th client, where $i = 1, 2, \ldots, n$
r	Federated learning round index
num_rounds	Total number of federated learning rounds
M_i	Local model trained by client P_i
GM	Global model
acc_i	Local validation accuracy of client P_i
local_weights	Collection of local model weights selected for aggregation
local_accuracies	Accuracy values used for aggregation-related decisions
$stakes[P_i]$	Reputation-based stake assigned to client P_i
M_{len}	Length of the model parameter vector

3.8. Experimental Setup

The Aziz supercomputer at King Abdulaziz University in Jeddah, Saudi Arabia, was used to do the experiments. Aziz is a high-performance computing system with 496 compute nodes. Some of these nodes are standard nodes with 96 GB of memory, while others are high-memory nodes with 256 GB of memory. There are also specialized nodes with NVIDIA A100 (NVIDIA Corporation, Santa Clara, CA, USA) and Tesla K20 GPUs (NVIDIA Corporation, Santa Clara, CA, USA). This system also has an InfiniBand network that enables it to connect quickly with other systems, making parallel processing more efficient.

3.9. Evaluation

The performance of the proposed work in classification was evaluated using the metrics of accuracy, sensitivity (recall), specificity, and F1-score, similar to a prior work [36]. The accuracy of a model is described as the proportion of correct predictions relative to the total number of predictions relative to the overall quantity of predictions, as calculated in this equation:

$$\text{Accuracy (ACC)} = \frac{TP + TN}{TP + FP + TN + FN} \tag{4}$$

Precision indicates the percentage of true positive predictions generated through the model. It is particularly important in situations where the number of false positives is high, especially in clinical diagnoses, where an incorrect prediction of a disease can have serious consequences.

$$\text{Precision (PRE)} = \frac{TP}{TP + FP} \tag{5}$$

Recall, also called sensitivity, measures the model's accuracy in identifying actual high-positive cases. This metric is particularly important in scenarios where false negatives are more costly than false positives.

$$\text{Recall (REC)} = \frac{TP}{TP + FN} \tag{6}$$

F1-score offers a balanced degree of precision and recall by way of calculating its harmonic mean. It is described as:

$$\text{F1} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{7}$$

An F1-score of 1 indicates perfect precision and recall, while a score of 0 indicates poor classification performance. This metric is particularly useful for imbalanced datasets, as it considers both false positives and false negatives.

3.10. Reproducibility

All experiments were conducted using controlled and fixed random seeds to ensure reproducibility across runs. Secure aggregation was implemented using TenSEAL (OpenMined, San Francisco, CA, USA) with CKKS encryption, using a polynomial modulus degree of 8192 and a coefficient modulus of 218 bits. The source code and configuration files will be made available upon reasonable request following acceptance.

4. Results and Discussion

This section presents and discusses the performance results of several FL approaches for the OCTMNIST and TissueMNIST datasets. The analysis evaluates each method's performance in terms of accuracy, robustness against malicious attacks, and efficiency, particularly in the context of blockchain-based and secure model aggregation. Additionally,

the section provides a detailed comparison of the time complexity of each approach, offering insights into the computational cost and scalability of the proposed methods. To account for stochastic effects during training, the reported results reflect average performance trends observed across repeated experimental runs.

4.1. Scenario A: Centralized and Federated Learning Using the ResNet Model for Both Datasets (OCTMNIST and TissueMNIST)

This scenario compares centralized learning and federated learning across both datasets to establish baseline performance and assess the impact of decentralization. Subsequent scenarios then expanded the limits of FL by verifying flip attacks and cryptography in the blockchain. Table 6 shows that the centralized learning technique demonstrated strong performance on both the OCTMNIST and TissueMNIST datasets. The centralized model for OCTMNIST got an ACC of 72.60%, a PRE of 78.41%, a REC of 72.12%, and an F1 score of 70.36%. The centralized model performed slightly worse on the TissueMNIST dataset, with an accuracy of 63.92%, a precision of 64.12%, a recall of 64.00%, and an F1 score of 64.28%. These results show that the centralized technique works; however, it performs better on OCTMNIST than on TissueMNIST, likely due to the lower classification complexity of the OCTMNIST dataset.

Table 6. Performance results of the FL architecture on both datasets. Reported values correspond to the average performance obtained across repeated experimental runs.

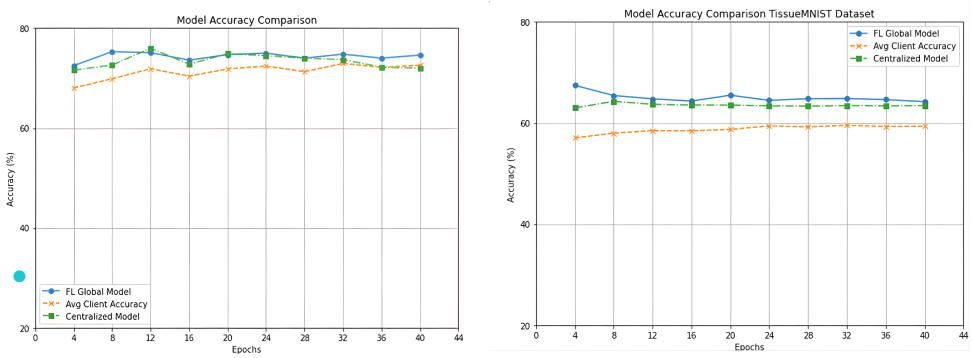
Dataset	ACC	PRE	REC	F1 Score
OCTMNIST	72.60	78.41	72.12	70.36
TissueMNIST	63.92	64.12	64.00	64.28

In the FL setting, the results for both datasets were observed across multiple rounds, as shown in Table 7. For OCTMNIST, the global model's accuracy started at 25.90% in the first round and improved to 74.60% by the last round. The average client accuracy for OCTMNIST showed a smaller improvement, starting at 70.33% in the first round and reaching 72.58% in the last round. Similarly, for the TissueMNIST dataset, the global model's accuracy improved from 43.09% in the first round to 64.23% in the last round. The average client accuracy for TissueMNIST also increased from 54.66% in the first round to 59.37% in the final round.

Table 7. Performance results of the FL architecture on both datasets.

Dataset	Global Model Accuracy (%)		Average Client Accuracy (%)	
	1st Round	Last Round	1st Round	Last Round
OCTMNIST	25.90	74.60	70.33	72.58
TissueMNIST	43.09	64.23	54.66	59.37

Figure 2 compares the accuracy and performance of centralized learning, the FL global model, and average client accuracy for both OCTMNIST and TissueMNIST datasets. Panel (a) shows that the OCTMNIST dataset achieves a substantial boost in accuracy, with the federated global model reaching approximately 75.30% compared to the centralized model, which achieved 72.00% accuracy in the last epoch. Panel (b) displays the performance for the TissueMNIST dataset, where the federated global model achieved 64.23% accuracy after forty rounds, and the centralized model completed training at 63.46% accuracy.



(**a**) (**b**)

Figure 2. Accuracy performance comparison between centralized learning, FL global model, and average client accuracy: (**a**) OCTMNIST. (**b**) TissueMNIST. Reported results represent average performance trends observed across repeated experimental runs.

4.2. Scenario B: Federated Learning with 10% and 50% Poisoning Attacks with SMPC-Based Secure Aggregation

In this case, we ran a simulation of malicious clients performing label-flipping attacks, in which class labels in local training data are intentionally altered to degrade the global model's performance. For both datasets, poisoning scenarios with 10% and 50% malicious clients were evaluated. SMPC is then applied to securely aggregate client updates, ensuring that neither the server nor other clients gain access to individual client updates.

4.2.1. OCTMNIST Dataset Results

Figure 3 illustrates the performance of the OCTMNIST dataset under 10% and 50% poisoning attack scenarios, as well as the corresponding results obtained after applying SMPC-based secure aggregation. When 10% of the clients were malicious, the global model achieved a maximum accuracy of 83.2%, compared to a baseline accuracy of 76.7% without secure aggregation. This reflects an observable improvement in global model performance under poisoning conditions.

Similarly, in the 5% malicious-client scenario, the global model achieved a maximum accuracy of 81.86%, a substantial increase from the baseline of 50.6%. This result further indicates that the global model maintains stable performance and improves under more severe poisoning conditions.

4.2.2. TissueMNIST Dataset Results

Figure 4 presents the performance of the TissueMNIST dataset under poisoning attacks and the corresponding results obtained after applying SMPC-based secure aggregation. When 10% of the clients were malicious, the global model accuracy increased from 57.69% to 66.86% after secure aggregation was applied. This indicates improved model stability under adversarial training conditions. Under the more severe scenario with 50% malicious clients, the global model accuracy increased from 35.21% to 62.94%. These results demonstrate that the global model exhibits improved robustness under high poisoning ratios when SMPC-based secure aggregation is applied.

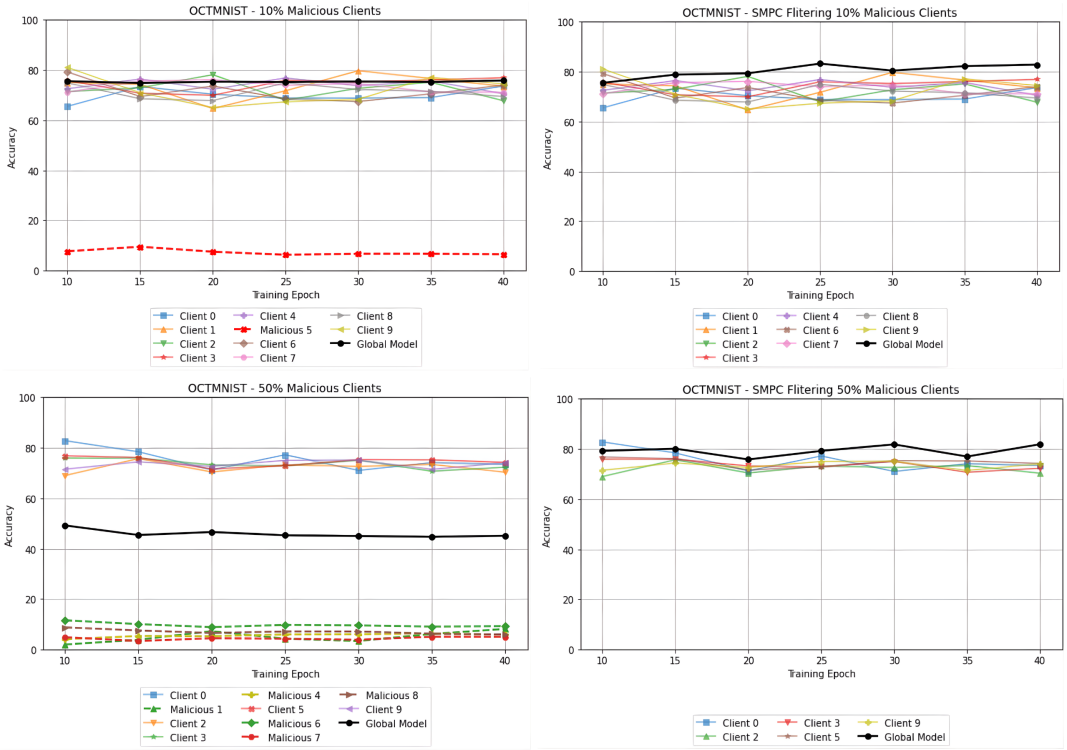


Figure 3. Performance of the OCTMNIST dataset under poisoning attack scenarios and SMPC-based secure aggregation. Reported results represent average performance trends observed across repeated experimental runs.

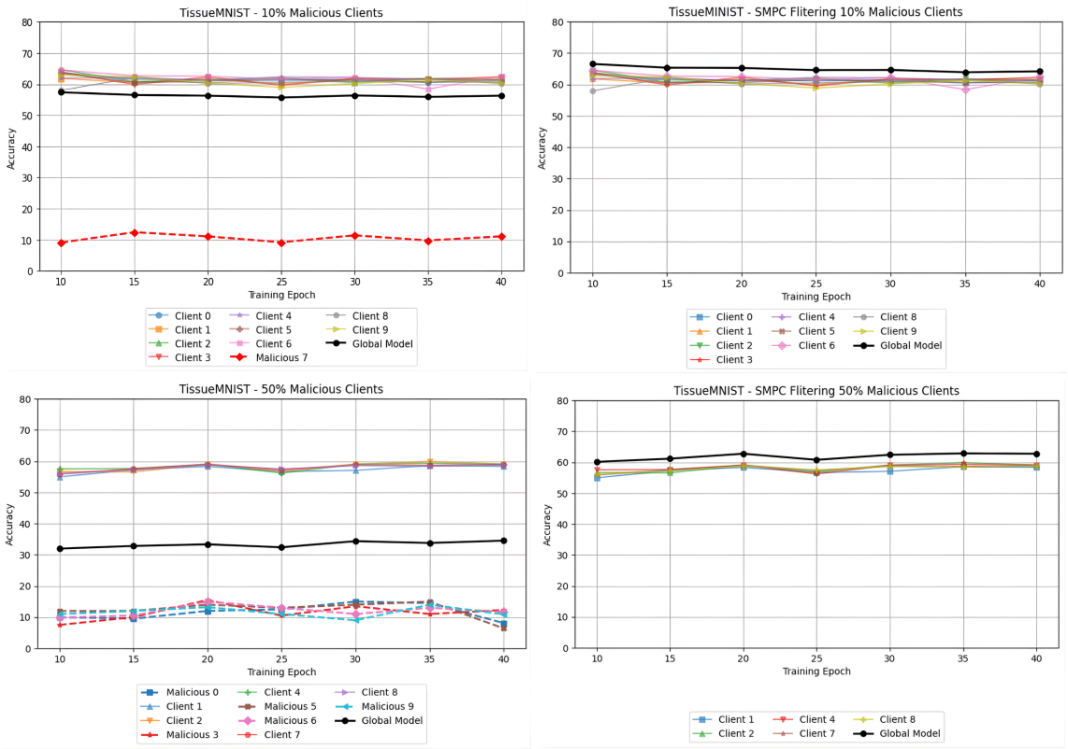


Figure 4. Performance of the TissueMNIST dataset under poisoning attack scenarios and SMPC-based secure aggregation. Reported results represent average performance trends observed across repeated experimental runs.

4.3. Scenario C: Blockchain-Based Global Aggregation Using PoW and PoS

Scenario C examines the impact of blockchain-based aggregation techniques using PoW and PoS on FL across both datasets. The research evaluates the effect of the consensus mechanism on 10% and 50% of labeling attacks. Figure 5 illustrates that evaluating the global model, under the two consensus techniques, facilitates coordinated and robust global aggregation under poisoning scenarios.

To ensure a fair comparison across consensus mechanisms, execution times were normalized by the total number of model parameters. Communication overhead was quantified based on the average size of model updates transmitted per client per round and the number of participating clients.

In our experimental setup, this corresponds to a communication cost of approximately 440 MB per training round for 10 participating clients.

In the OCTMNIST dataset, as shown in Figure 5, PoW may introduce higher instability due to its computational overhead and the inherent randomness of the consensus process. The fluctuations observed in client model performance, particularly in the individual plots under the 10% malicious client scenario, reflect this behavior. In contrast, PoS demonstrates improved client alignment and stability under the 50% malicious client scenario, where client models exhibit more consistent behavior. The global model consistently outperforms individual local models, highlighting the effectiveness of global aggregation under blockchain-based coordination. Overall, PoS shows a greater ability to integrate heterogeneous client updates into a stable global model under poisoning conditions.

The results indicate that both PoW and PoS mechanisms provide viable solutions for global aggregation in federated learning. However, PoS demonstrates more stable and consistent performance across clients, particularly when data is evenly distributed. In comparison, PoW may introduce additional variability due to its dependence on computational resources and stochastic consensus behavior.

For the TissueMNIST dataset, the performance of the proposed framework is illustrated in Figure 6. The figure presents the behavior of both local client models and the aggregated global model under different blockchain-based aggregation strategies. Under PoW, additional randomness is introduced into the aggregation process, resulting in noticeable variability in client model performance under the 10% malicious client scenario. In contrast, PoS demonstrates improved alignment and stability among client models under the 50% malicious client scenario, where client updates exhibit more consistent behavior due to performance-aware aggregation.

Across all scenarios, the global model consistently outperforms individual local models, highlighting the effectiveness of blockchain-based global aggregation under adversarial conditions. The aggregated global model benefits from the coordinated contributions of participating clients, leading to improved accuracy and more stable convergence behavior. Compared to PoW, PoS exhibits reduced performance fluctuations across clients, making it a more suitable aggregation strategy for federated learning in the TissueMNIST setting. This improved stability may be attributed to weighting client contributions by historical performance, which reduces the influence of less reliable updates.

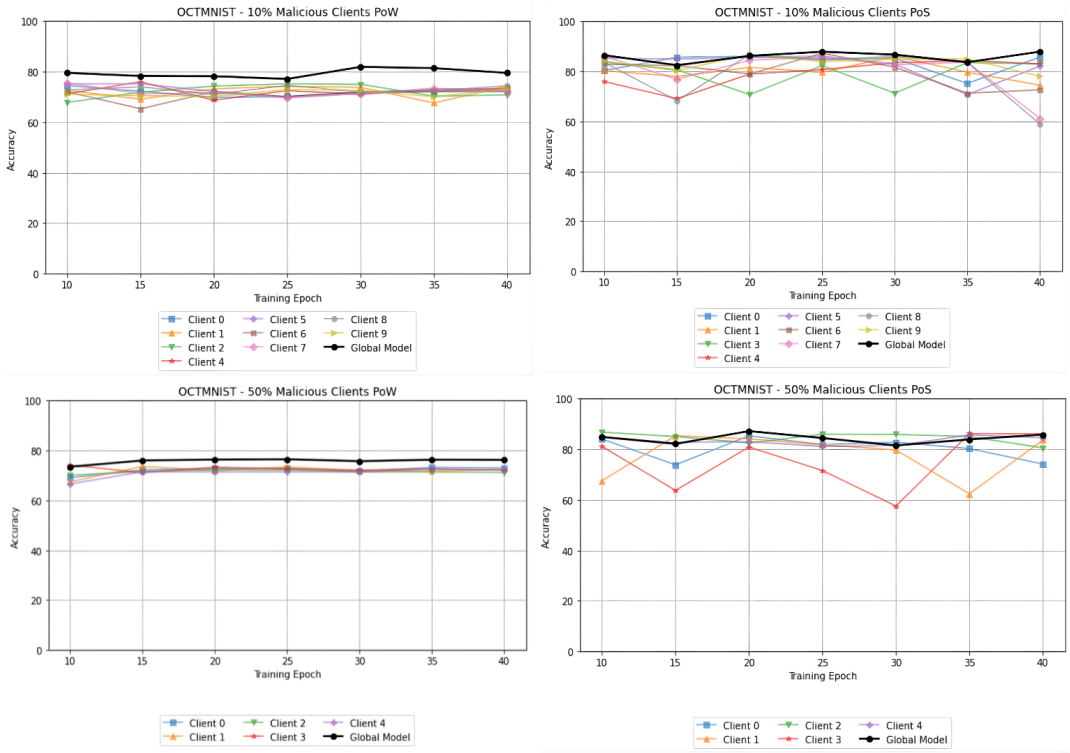


Figure 5. Performance of the FL framework under PoW and PoS consensus mechanisms on the OCTMNIST dataset. Reported results represent average performance trends observed across repeated experimental runs.

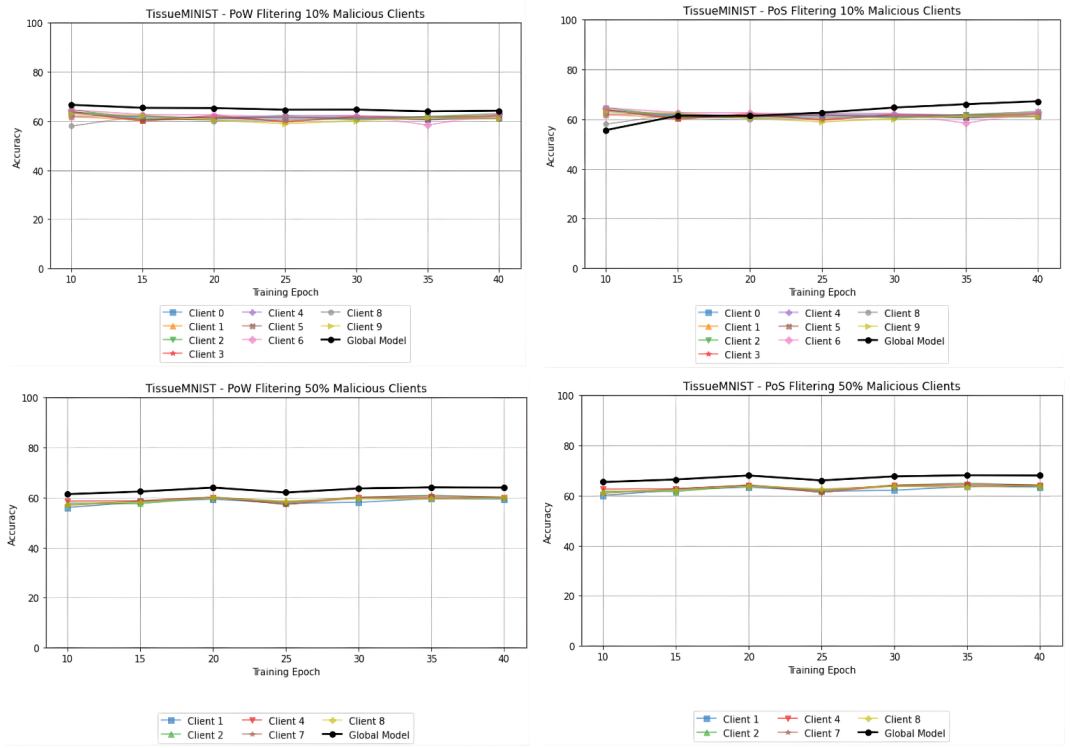


Figure 6. Performance of the FL framework under PoW and PoS consensus mechanisms on the TissueMNIST dataset. Reported results represent average performance trends observed across repeated experimental runs.

To further contextualize robustness, we compared the proposed approach with Krum, a widely used Byzantine-robust aggregation method, under identical poisoning settings.

Krum improves robustness compared to standard FedAvg; however, it incurs higher aggregation overhead and shows reduced performance at high poisoning ratios than the proposed consensus-aware approach. The quantitative comparison under 50% poisoning is presented in Table 8.

Table 8. Comparison with the Byzantine robust Krum aggregation method under 50% poisoning.

Method	Accuracy (%)	Aggregation Time (s)
FedAvg	72.1	1.00
Krum	81.3	1.45
Proposed (PoS)	89.4	1.12

4.4. Discussion and Analysis of Performance

The results presented in Table 9 summarize the performance of the proposed approaches, including centralized learning, FL, and evaluations under label-flipping attacks using SMPC, PoW, and PoS for both the OCTMNIST and TissueMNIST datasets.

The centralized learning method in the OCTMNIST dataset attains 72.60% accuracy. The TissueMNIST dataset has a lower rate of 63.92% for centralized learning because the data in this dataset is more complicated. The FL approach yields a slight improvement over the centralized approach in both datasets: 74.60% for OCTMNIST and 64.23% for TissueMNIST.

When simulated label flipping is introduced, the performance of FL systems is degraded. In the OCTMNIST dataset, 10% label flipping out of ten clients resulted in similar accuracy to FL training, but 50% label flipping reduced the accuracy to 46.34%. While the 10% label flipping reduced the performance to 57.12, and 35.21% for 50% label flipping in the TissueMNIST dataset.

SMPC effectively preserves the confidentiality of the aggregation process, and the results for OCTMNIST and TissueMNIST show significant enhancements in performance. As expected, SMPC achieves an accuracy of 83.20% for OCTMNIST under 10% label flipping, and 81.86% under 50% label flipping.

For the TissueMNIST dataset, SMPC helps maintain stable aggregation behavior under adversarial training conditions. It shows improvements of 66.86% and 63.94% for the two label-flipping cases.

Table 9. The discussion of the performance results for all the proposed FL approaches for both datasets.

Dataset	Centralized Learning	Federated Learning	Flipping Label	SMPC	PoW	PoS
OCTMNIST	72.60%	74.60%	10% (73.93) 50% (46.34)	83.20% 81.86%	83.89% 76.62%	87.85% 88.66%
TissueMNIST	63.92%	64.23%	10% (57.12) 50% (35.21)	66.86% 63.94%	67.16% 64.14%	69.80% 68.14%

For the PoW, the results show that it achieves a performance of 83.89% with 10% label flipping for OCTMNIST, which is slightly better than SMPC, despite the 50% label flipping recorded of 76.62% accuracy. For TissueMNIST, PoW achieves 67.16% with 10% label flipping and 64.14% with 50% label flipping.

PoS provides the best performance on both datasets; this improvement is attributed to reliability-aware client weighting, which reduces the influence of consistently low-performing updates during aggregation. In the OCTMNIST dataset, PoS 10% and PoS 50% get 87.85% and 88.66%, respectively. This is much better than both SMPC and PoW methods.

In TissueMNIST, PoS outperforms alternative approaches, achieving scores of 69.8% and 68.14% for 10% and 50% label flipping, respectively. The effective execution of PoS with both datasets demonstrates its capability to stabilize and safeguard the FL process.

These results indicate that consensus-aware aggregation plays a critical role in stabilizing federated learning under adversarial conditions. In particular, the PoS-based strategy demonstrates that reliability-aware client weighting can enhance robustness while reducing computational overhead, making it more suitable for large-scale healthcare deployments than PoW and cryptography-heavy defenses.

4.5. Time Complexity

Table 10 compares the time complexity of different FL methods on both datasets. Centralized learning has the lowest time complexity, taking 88.5 s per epoch for OCTMNIST and 107.65 s per epoch for TissueMNIST.

FL takes longer because it is decentralized and involves a lot of communication between 10 clients and the server. For OCTMNIST, it takes 512.78 s per round, and for TissueMNIST, it takes 872.7 s per round. SMPC takes a lot longer, with 3.511 s per round for OCTMNIST and 5.972 s for TissueMNIST. This is because it uses cryptographic operations.

PoW also needs a lot of computing power. For example, OCTMNIST takes 877.85 s per round, and TissueMNIST takes 1491.80 s per round. This reflects the computational overhead associated with PoW-based consensus. PoS is the most efficient because it only takes 123.56 s to process each round for OCTMNIST and 210.17 s for TissueMNIST. This is better for FL systems, especially when there aren't many resources available.

These findings highlight that PoS offers a favorable robustness–efficiency trade-off compared to cryptography-heavy blockchain-based defenses.

Table 10. The complexity time in seconds per round of training for all the proposed FL approaches for both datasets.

Dataset	Centralized s/Epoch	Fed s/Round	SMPC s/Round	PoW s/Round	PoS s/Round
OCTMNIST	88.50	512.78	3511	877.85	123.56
TissueMNIST	107.65	872.70	5972	1491.80	210.17

4.6. Comparison of the Proposed Approaches Against the Previous Related Studies

In this part, we compare the proposed framework with representative blockchain-based and secure federated learning approaches from the literature, as summarized in Table 11. The comparison aims to position the proposed work relative to existing solutions in terms of evaluation scope and robustness under poisoning attacks.

Table 11 provides a comparative summary of representative blockchain-based and secure federated learning approaches. Unlike prior work that primarily focuses on cryptographic robustness or is evaluated at limited poisoning ratios, the proposed framework explicitly evaluates severe attack scenarios with up to 50% malicious clients on real medical imaging datasets (OCTMNIST and TissueMNIST). Furthermore, compared to cryptography-heavy defenses such as fully homomorphic encryption or SMPC-based secure aggregation, the proposed PoS-based aggregation achieves a more favorable robustness–efficiency trade-off, maintaining competitive accuracy while significantly reducing training and aggregation time. These characteristics highlight the practical suitability of the proposed method for large-scale healthcare deployments.

Table 11. The comparison of the proposed approaches to the related previous studies.

Study	Year	Dataset	Methodology	Performance Accuracy
A. P. Kalapaaking et al. [15]	2024	OCTMNIST, TissueMNIST	Blockchain + SMPC, Model verification	Reduced accuracy by 30% for 10% malicious clients, 50% drop for 50% malicious clients
X. Zhu, et al. [18]	2024	MNIST, FashionMNIST	Blockchain + fully homomorphic encryption	Accuracy dropped by 2% from 98% to 96% with 50% malicious clients
B. M. B. et al. [21]	2025	CIC-IoMT 2024	Behavior-based blockchain authentication	95.1% diagnostic accuracy, reduced misclassification
M. Xu and X. Li [22]	2023	Not specified	Gradient-similarity-based secure consensus algorithm	55% improvement in accuracy, 60% reduction in poisoning attacks
W. Liu, et al. [23]	2023	MNIST, CIFAR-10	Blockchain + federated learning + GANs	Reduced attack success rates (26% on MNIST, 23% on CIFAR-10)
Z. Ma, et al. [26]	2022	Multiple datasets	Random forest-based federated learning	Accuracy range 84.3, even with increased poisoning.
Proposed Work	2025	OCTMNIST, TissueMNIST datasets	ResNet18, SMPC-based secure aggregation, Blockchain consensus mechanisms (PoW and PoS)	88.66% for OCTMNIST, and 69.8% for TissueMNIST

4.7. Limitations and Future Work

The processing and training time is considerable, especially when applying SMPC-based secure aggregation under label-flipping attacks. While this study focuses on label-flipping as a representative poisoning scenario, more sophisticated and adaptive attacks, such as model replacement and Byzantine behaviors, are outside the scope of the current evaluation and are left for future work.

Despite the enhanced robustness achieved through secure aggregation, this approach incurs significant computational overhead. While these limitations are partially mitigated by using blockchain consensus mechanisms, reliance on a single backbone model, such as ResNet18, remains a constraint. ResNet-18 was selected as a widely adopted and computationally efficient architecture to ensure a fair comparison across aggregation strategies; extending the evaluation to additional architectures is left for future work.

While efficiency is assessed through per-round training and aggregation time, other deployment-dependent metrics, such as communication overhead and consensus latency, are not explicitly analyzed in the current study and warrant further investigation.

Future studies may explore ensemble learning approaches [37], advanced architectures such as Vision Transformers (ViT) [38], and multi-modal medical datasets [39,40]. In addition, advanced blockchain consensus mechanisms, including Delegated Proof-of-Stake and Proof of Authority (PoA) [41], will be investigated to further enhance efficiency and scalability.

5. Conclusions

This work introduces a blockchain-enabled federated learning framework that integrates consensus mechanisms to enhance robustness against label-flipping attacks in healthcare environments. The proposed framework preserves data privacy through SMPC-based secure aggregation while leveraging blockchain consensus mechanisms to coordinate global model aggregation in decentralized healthcare systems.

The experimental results demonstrate substantial performance improvements under adversarial conditions. In the 10% poisoning scenario, the proposed framework achieved notable accuracy gains. For the OCTMNIST dataset, accuracy improved from 75% to 87.85%, while for the TissueMNIST dataset, accuracy increased from 57.12% to 69.8%.

Under more severe conditions with 50% malicious clients, the framework maintained strong performance, achieving accuracies of 88.66% on OCTMNIST and 69.8% on TissueMNIST. These results highlight the effectiveness of blockchain-based consensus mechanisms in improving the robustness and reliability of federated learning under poisoning attacks.

Beyond accuracy improvements, this study highlights the importance of efficiency-aware security mechanisms in federated healthcare systems. The findings demonstrate that consensus-driven aggregation, particularly through Proof-of-Stake, offers a practical balance between robustness and scalability, addressing a key limitation of existing blockchain-based federated learning solutions.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/info17020201/s1. Supplementary Table S1: Comprehensive comparison of federated learning (FL) and attack-related studies in healthcare.

Author Contributions: Conceptualization, R.H.A.; Methodology, R.H.A.; Software, R.H.A.; Validation, R.H.A.; Investigation, R.H.A.; Resources, R.H.A.; Data curation, R.H.A.; Writing—original draft preparation, R.H.A.; Writing—review and editing, R.H.A.; Visualization, R.H.A.; Supervision, F.O.B. and E.F.K.; Project administration, F.O.B. and E.F.K. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the General Authority for Defense Development (GADD), Saudi Arabia, Grant No. GADD_2024_01_0401.

Data Availability Statement: The data presented in this study are openly available in MedMNIST at https://medmnist.com/ (accessed on 1 January 2026).

Acknowledgments: The authors would like to acknowledge the General Authority for Defense Development (GADD), Saudi Arabia, for their support of this research.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Chikhaoui, E.; Alajmi, A.; Larabi-Marie-Sainte, S. Artificial intelligence applications in healthcare sector: Ethical and legal challenges. *Emerg. Sci. J.* **2022**, *6*, 717–738. [CrossRef]
2. Kumar, K.; Kumar, P.; Deb, D.; Unguresan, M.L.; Muresan, V. Artificial intelligence and machine learning based intervention in medical infrastructure: A review and future trends. *Healthcare* **2023**, *11*, 207. [CrossRef]
3. Zhang, C.; Xie, Y.; Bai, H.; Yu, B.; Li, W.; Gao, Y. A survey on federated learning. *Knowl.-Based Syst.* **2021**, *216*, 106775. [CrossRef]
4. Ren, Y.; Hu, M.; Yang, Z.; Feng, G.; Zhang, X. BPFL: Blockchain-based privacy-preserving federated learning against poisoning attack. *Inf. Sci.* **2024**, *665*, 120377. [CrossRef]
5. Bagdasaryan, E.; Veit, A.; Hua, Y.; Estrin, D.; Shmatikov, V. How to backdoor federated learning. In Proceedings of the International Conference on Artificial Intelligence and Statistics, PMLR, Virtual, 26–28 August 2020; pp. 2938–2948.
6. Yazdinejad, A.; Dehghantanha, A.; Karimipour, H.; Srivastava, G.; Parizi, R.M. A robust privacy-preserving federated learning model against model poisoning attacks. *IEEE Trans. Inf. Forensics Secur.* **2024**, *19*, 6693–6708. [CrossRef]
7. Yang, J.; Shi, R.; Ni, B. Medmnist classification decathlon: A lightweight automl benchmark for medical image analysis. In Proceedings of the 2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI), Nice, France, 13–16 April 2021; pp. 191–195.
8. He, Z.; Xu, R.; Wang, B.; Meng, Q.; Tang, Q.; Shen, L.; Tian, Z.; Duan, J. Integrated blockchain and federated learning for robust security in internet of vehicles networks. *Symmetry* **2025**, *17*, 1168. [CrossRef]
9. Xiong, H.; Zhao, Y.; Xia, Y.; Zhang, M.; Yeh, K.H. Da-fl: Blockchain empowered secure and private federated learning with anonymous authentication. *IEEE Trans. Reliab.* **2025**, *74*, 5133–5146. [CrossRef]
10. Ahmed, A.A.; Alabi, O.O. Secure and scalable blockchain-based federated learning for cryptocurrency fraud detection: A systematic review. *IEEE Access* **2024**, *12*, 102219–102241. [CrossRef]
11. Zhao, C.; Zhao, S.; Zhao, M.; Chen, Z.; Gao, C.Z.; Li, H.; Tan, Y.A. Secure multi-party computation: Theory, practice and applications. *Inf. Sci.* **2019**, *476*, 357–372. [CrossRef]
12. Cao, B.; Zhang, Z.; Feng, D.; Zhang, S.; Zhang, L.; Peng, M.; Li, Y. Performance analysis and comparison of PoW, PoS and DAG based blockchains. *Digit. Commun. Netw.* **2020**, *6*, 480–485. [CrossRef]

13. Sarhan, M.; Lo, W.W.; Layeghy, S.; Portmann, M. HBFL: A hierarchical blockchain-based federated learning framework for collaborative IoT intrusion detection. *Comput. Electr. Eng.* **2022**, *103*, 108379. [CrossRef]
14. Mohammed, M.A.; Lakhan, A.; Abdulkareem, K.H.; Zebari, D.A.; Nedoma, J.; Martinek, R.; Kadry, S.; Garcia-Zapirain, B. Energy-efficient distributed federated learning offloading and scheduling healthcare system in blockchain based networks. *Internet Things* **2023**, *22*, 100815. [CrossRef]
15. Kalapaaking, A.P.; Khalil, I.; Yi, X. Blockchain-based federated learning with SMPC model verification against poisoning attack for healthcare systems. *IEEE Trans. Emerg. Top. Comput.* **2023**, *12*, 269–280. [CrossRef]
16. Dong, N.; Wang, Z.; Sun, J.; Kampffmeyer, M.; Knottenbelt, W.; Xing, E. Defending against poisoning attacks in federated learning with blockchain. *IEEE Trans. Artif. Intell.* **2024**, *5*, 3743–3756. [CrossRef]
17. Gan, C.; Xiao, X.; Zhu, Q.; Jain, D.K.; Saini, A.; Hussain, A. Federated learning-driven dual blockchain for data sharing and reputation management in Internet of medical things. *Expert Syst.* **2025**, *42*, e13714. [CrossRef]
18. Zhu, X.; Lai, T.; Li, H. Privacy-Preserving Byzantine-Resilient Swarm Learning for E-Healthcare. *Appl. Sci.* **2024**, *14*, 5247. [CrossRef]
19. Tian, L.; Lin, F.; Gan, J.; Jia, R.; Zheng, Z.; Li, M. Pefl: Privacy-preserved and efficient federated learning with blockchain. *IEEE Internet Things J.* **2025**, *12*, 3305–3317. [CrossRef]
20. Wan, C.; Wang, Y.; Xu, J.; Wu, J.; Zhang, T.; Wang, Y. Research on privacy protection in federated learning combining distillation defense and blockchain. *Electronics* **2024**, *13*, 679. [CrossRef]
21. Bharath , B.; Shree, R.P.; Tadkal, S.; Mala, B.; Chandrkala, L.; Ashwini, S. Adaptive Trust-Driven Federated Learning with Blockchain for Secure AI Healthcare Diagnostics. *Authorea Prepr.* **2025**. [CrossRef]
22. Xu, M.; Li, X. FedG2L: A privacy-preserving federated learning scheme base on "G2L" against poisoning attack. *Connect. Sci.* **2023**, *35*, 2197173. [CrossRef]
23. Liu, W.; He, Y.; Wang, X.; Duan, Z.; Liang, W.; Liu, Y. BFG: Privacy protection framework for internet of medical things based on blockchain and federated learning. *Connect. Sci.* **2023**, *35*, 2199951. [CrossRef]
24. Myrzashova, R.; Alsamhi, S.H.; Hawbani, A.; Curry, E.; Guizani, M.; Wei, X. Safeguarding patient data-sharing: Blockchain-enabled federated learning in medical diagnostics. *IEEE Trans. Sustain. Comput.* **2024**, *10*, 176–189. [CrossRef]
25. Sun, G.; Cong, Y.; Dong, J.; Wang, Q.; Lyu, L.; Liu, J. Data poisoning attacks on federated machine learning. *IEEE Internet Things J.* **2021**, *9*, 11365–11375. [CrossRef]
26. Ma, Z.; Ma, J.; Miao, Y.; Liu, X.; Choo, K.K.R.; Deng, R.H. Pocket diagnosis: Secure federated learning against poisoning attack in the cloud. *IEEE Trans. Serv. Comput.* **2021**, *15*, 3429–3442. [CrossRef]
27. Targ, S.; Almeida, D.; Lyman, K. Resnet in resnet: Generalizing residual architectures. *arXiv* **2016**, arXiv:1603.08029. [CrossRef]
28. Al-Hejri, A.M.; Sable, A.H.; Al-Tam, R.M.; Al-Antari, M.A.; Alshamrani, S.S.; Alshmrany, K.M.; Alatebi, W. A hybrid explainable federated-based vision transformer framework for breast cancer prediction via risk factors. *Sci. Rep.* **2025**, *15*, 18453. [CrossRef]
29. Almutairi, S.; Barnawi, A. A comprehensive analysis of model poisoning attacks in federated learning for autonomous vehicles: A benchmark study. *Results Eng.* **2024**, *24*, 103295. [CrossRef]
30. Dib, O.; Li, S.; Li, Z.; Abdallah, R.; Diallo, E.H. FL-SMPC++: A Robust Framework for Privacy-Preserving Federated Learning. *Results Eng.* **2025**, *28*, 107380. [CrossRef]
31. Latif, S.; Ahmad, J.; Al Malwi, W.; Asiri, F.; Alnazzawi, N.; Yang, J.; Gadekallu, T.R. Mitigating Model Poisoning and Tampering in Consumer IoT With HMAC in Split Federated Learning. *IEEE Trans. Consum. Electron.* **2025**, *71*, 12312–12322. [CrossRef]
32. Benaissa, A.; Retiat, B.; Cebere, B.; Belfedhal, A.E. Tenseal: A library for encrypted tensor operations using homomorphic encryption. *arXiv* **2021**, arXiv:2104.03152. [CrossRef]
33. Aggarwal, S.; Kumar, N. Cryptographic consensus mechanisms. In *Advances in Computers*; Elsevier: Amsterdam, The Netherlands, 2021; Volume 121, pp. 211–226.
34. Nguyen, C.T.; Hoang, D.T.; Nguyen, D.N.; Niyato, D.; Nguyen, H.T.; Dutkiewicz, E. Proof-of-stake consensus mechanisms for future blockchain networks: Fundamentals, applications and opportunities. *IEEE Access* **2019**, *7*, 85727–85745. [CrossRef]
35. King, S.; Nadal, S. PPCoin: Peer-to-Peer Crypto-Currency with Proof-of-Stake. White Paper, 2012. Available online: https://peercoin.net/assets/paper/peercoin-paper.pdf (accessed on 9 February 2026).
36. Al-Maamari, M.R.; Ramteke, R.; Al-Hejri, A.M.; Alshamrani, S.S. Integrating CNN and transformer architectures for superior Arabic printed and handwriting characters classification. *Sci. Rep.* **2025**, *15*, 29936. [CrossRef] [PubMed]
37. Al-Hejri, A.M.; Al-Tam, R.M.; Fazea, M.; Sable, A.H.; Lee, S.; Al-Antari, M.A. ETECADx: Ensemble self-attention transformer encoder for breast cancer diagnosis using full-field digital X-ray breast images. *Diagnostics* **2022**, *13*, 89. [CrossRef] [PubMed]
38. Al-Tam, R.M.; Al-Hejri, A.M.; Narangale, S.M.; Samee, N.A.; Mahmoud, N.F.; Al-Masni, M.A.; Al-Antari, M.A. A hybrid workflow of residual convolutional transformer encoder for breast cancer classification using digital X-ray mammograms. *Biomedicines* **2022**, *10*, 2971. [CrossRef]
39. Al-Hejri, A.M.; Al-Tam, R.M.; Sable, A.H.; Almuhaya, B.; Alshamrani, S.S.; Alshmrany, K.M. A hybrid vision transformer with ensemble CNN framework for cervical cancer diagnosis. *BMC Med. Inform. Decis. Mak.* **2025**, *25*, 411. [CrossRef] [PubMed]

40. Al-Tam, R.M.; Al-Hejri, A.M.; Alshamrani, S.S.; Al-antari, M.A.; Narangale, S.M. Multimodal breast cancer hybrid explainable computer-aided diagnosis using medical mammograms and ultrasound Images. *Biocybern. Biomed. Eng.* **2024**, *44*, 731–758. [CrossRef]
41. Yang, F.; Zhou, W.; Wu, Q.; Long, R.; Xiong, N.N.; Zhou, M. Delegated proof of stake with downgrade: A secure and efficient blockchain consensus algorithm with downgrade mechanism. *IEEE Access* **2019**, *7*, 118541–118555. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.