IJEMD-CSAI, Special Issue (2026) https://doi.org/ 10.54938/ijemdcsai.2026.04.2.608

IJEMD

International Journal of Emerging Multidisciplinaries:

Computer Science and Artificial Intelligence

Research Paper

Journal Homepage: www.ojs.ijemd.com

ISSN (print): 2791-0164 ISSN (online): 2957-5036

Towards Privacy-Preserving And Explainable CNN-Based Image Classification:

A Federated Learning and XAI Framework Grounded in Fashion-MNIST Engineering

Deng Mile [1], Abdul Salam Shah Sayed [1]

1. School of Computer Science, Taylor's University, Malaysia.

ABSTRACT

The paper introduces a two-axis extension of the standard Convolutional Neural Network (CNN) image classification studies; the technical exploration is based on an experimental baseline of Fashion-MNIST and generalizes its results to two frontier research directions of pressing societal importance Federated Learning (FL) to train distributed models privately and Explainable Artificial Intelligence (XAI) to make transparent and clinically interpretable decisions. It is a three-block sequential CNN (with 32, 64, and 64 filters Convolutional layers, max-pooling, dense classification and softmax output) that is trained on the 15,000 sample Fashion-MNIST test data with Adam optimizer categorical cross-entropy in 15 epochs (resulting in 89.57 percent accuracy and macro-averaged F1-score of 0.90). This performance profile per-class, i.e., high F1 on morphologically distinct classes (Trouser: 0.98; Bag: 0.98; Sandal: 0.96) and significantly lower performance on visually confusable classes (Shirt: 0.71) is systematically studied to incentivise a Federated Learning architecture that can quickly train CNNs on decentralised and non-IID data partitions without access to raw training examples and a Gradient-weighted Class Activation Mapping (Grad-CAM) XAI addition that can make Combined, these extensions map out a technically sound research path to deploy privacy-conscious, transparent CNN classifiers in high-stakes areas of application such as clinical diagnostic imaging, federated retail AI, and image pathology across institutions.

Keywords: Convolutional Neural Networks, Fashion-MNIST, Federated Learning, Explainable AI, Grad-CAM, Privacy-Preserving Deep Learning, Adam Optimizer, Categorical Cross-Entropy, Non-IID Data, Clinical AI.

1. INTRODUCTION

The ability of deep learning models to reason in high-dimensional visuals to even higher accuracy than humans has made Convolutional Neural Networks the architecture of applied AI to be utilized across the spectrum of applications in e-commerce and autonomous navigation, to clinical diagnostics and security monitoring. However, the extensive use of CNN-based classifiers is challenged by two structural issues, which the conventional benchmark evaluation paradigm fails to consider: privacy of the data trained on it, and the interpretability of decisions made by it.

On privacy: to train high-performing CNNs, the datasets need to be large, diverse, and accurately labeled - which in sensitive areas of our society, like healthcare, finance and legal record processing is directly in conflict with regulations that govern the utilization of personal data. The General Data Protection Regulation (GDPR, 2016) and [11] place rigorous limitations on aggregation of data in centralised data to train AI. Specifically in healthcare, more than 30 percent of organisations all around the world suffered at least one data breach in 2023 [15], and the U.S. Health Insurance Portability and Accountability Act (HIPAA) does not allow the centralised aggregation of patient imaging data across organisations without explicit consent. These issues render the conventional template of centralised data gathering and thereafter centralised CNN teaching logically impracticable to the most valuable and uncommon collection of training information.

Regarding interpretability: a CNN with a classification accuracy of 89.57 on a benchmark test set itself gives a human reviewer, such as a clinician or a regulatory auditor or an affected individual, no reason to believe that the decisions made by the CNN are motivated by semantically significant, non-discriminatory features of the data, as opposed to random correlations to the artifacts of the dataset. The requirements of explainability of high-risk AI systems are stipulated in [11]; the FDA guidance on AI/ML-based Software as a Medical Device (SaMD, 2021) also outlines that the advent of algorithm use must be a subject of understanding to the qualified medical professional. Accurate blackbox CNN classifiers, however, cannot meet these demands without the addition of either post-hoc or architecture-native explainability mechanisms.

This paper discusses these two barriers as a form of research agenda. Section 2 conducts a literature review. Section 3 describes the implementation of the Fashion-MNIST CNN baseline. The experimental findings are provided and analysed in Section 4. The Federated Learning extension framework is developed in Section 5. Section 6 comes up with the architecture of XAI/Grad-CAM integration. Section 7 indicates the most fruitful future research directions at the cross-sections of these three areas.

2. LITERATURE REVIEW

2.1 CNN Architectural Foundations.

Modern CNN Modern CNN dates back to the LeNet-5 [23], which introduced the rule that banks of learned convolutional filters moving over image spatial dimensions with shared weights are effective at capturing translation-invariant local features on which pattern recognition can be done. This was later scaled to the 1.2-million-image ImageNet benchmark by AlexNet [22] with a relative error rate of 10.8% by using ReLU activation, dropout regularization [37] and training on a GPU. VGGNet [35] proved that the first factor in representational capacity is depth which is the result of uniform convolution stacking of 3x3 and ResNet (He et al., 2016) address the vanishing gradient problem in extremely deep networks by using identity skip connections. All further applications of CNN domains are based on these architectural innovations.

CNNs are more effective on Fashion-MNIST in particular, with the best results of SIFT+SVM, about 82-85%, and CNNs, ranging between 88-93 with good settings, regularization, and augmentation choice [42]; [28]. [1] also illustrated that per-class performances of 90.6-100% on multi-class visual benchmarks can be obtained by even lightweight custom CNNs when trained using TensorFlow, confirming the feasibility of from-scratch CNN training in applied classification tasks.

2.2 Federated CNN Training Federated Learning.

[26] introduced Federated Learning (FL) as a communicatively efficient distributed learning paradigm where a global model is optimized by sequential combination of locally computed gradient updates no raw data is ever taken out of the device or the institution that created it. During every FL round, there is a central aggregation server that gives out the current global model weights to the engaging clients; clients execute local SGD on their own data partition and submit only the updated model gradients; the server gathers them with FedAvg or an equivalent and updates the global model. The privacy of local data is ensured by design with this architecture: no training samples are ever sent, and even model gradients (which contain some statistical information about the local data) can be privatized using the methods of differential privacy (DP): gradient clipping and gradient noise injection [9].

In medical imaging, FL has demonstrated particular utility. [31] surveyed FL applications across radiology, pathology, and ophthalmology, reporting that federated models trained across multiple hospitals consistently approach the accuracy of centrally trained models while fully preserving patient data locality. A 2025 Nature Scientific Reports study [41] achieved high classification accuracy across TB chest X-rays, brain tumour MRI, and diabetic retinopathy datasets using GoogLeNet and VGG16 in a federated transfer learning framework, explicitly addressing the data privacy challenge that prevents centralised training in clinical settings. [39] conducted a systematic review of federated machine learning in healthcare, finding that federated CNN models match or approach centrally trained baselines in 78% of surveyed studies across diverse imaging modalities, validating FL as a clinically viable training paradigm.

2.3 Explainable AI for CNN Decision Transparency

The non-transparency of deep CNN classifiers - that is, their black box character where high-dimensional weight matrices are used to convert pixels on the input into class probabilities by a host of nonlinear functions without any human interpretable intermediate representations - is a fundamental impediment to their use in decision systems with high stakes. Explainable Artificial Intelligence (XAI) overcomes this obstacle by methods that assign model predictions to particular features of input to make post-hoc explanations of individual classification outputs. The most commonly used spatial XAI technique of CNN classifiers is Grad-CAM (Gradient-weighted Class Activation Mapping; Selvaraju et al., 2017). It calculates the gradient of the classification score with respect to the feature maps of the last convolutional layer, weights every feature map by the mean gradient magnitude of that feature map across the whole image, and results in a spatial heatmap of the parts of the image most important to the decision made by the model. When applied to Fashion-MNIST, Grad-CAM would show that the bifurcated lower silhouette area is influencing the Trouser classification of the model, and the Shirt / T-shirt confusion would be reflected in overlapping patterns of activations in the collar and shoulders areas - diagnostic information that can be acted upon to refine the architecture. Grad-CAM heatmaps on top of a chest X-ray [29], brain MRI image (Frontiers AI, 2025), and dermoscopy image (Esteva et al., 2017) have been demonstrated to produce clinically valid attention localizations, confirmed by radiologists. A 2024 PMC study comparing Grad-CAM and LIME to pneumonia and COVID-19 that examined how well they did on clinical relevance and comprehensibility identified that 26 medical experts rated the Grad-CAM explanations positively but stated that there was little awareness of the XAI tools in the current radiological workflow [20].

Table 1: Benchmark Comparison — CNN Architectures on Fashion-MNIST and Medical Datasets

Model / Study	Domain	Dataset	Accuracy / F1	Key Innovation
LeNet-5 (LeCun et al., 1998)	CV Benchmark	MNIST	99.2% Acc.	Shared weights, pooling
AlexNet (Krizhevsky et al., 2012)	CV Benchmark	ImageNet	63.3% Top-5	ReLU, Dropout, GPU
Nocentini et al. (2022)	CV Benchmark	Fashion-MNIST	92.1% Acc.	Multiple CNN layers
This Work (Deng Mile, 2026)	CV Benchmark	Fashion-MNIST	89.57% / F1=0.90	Sequential 3-block CNN
McMahan et al. (2017) — FL	Federated	CIFAR-10	~85% (federated)	FedAvg aggregation
Williamson & Prybutok (2025)	Medical + FL	TB, Brain MRI, DR	High multi-task Acc.	FL + Transfer Learning
Rajpurkar et al. (2017) — CheXNet	Medical XAI	ChestX-ray14	F1=0.435 (>radiologist)	DenseNet-121 + CAM
Kassem et al. (2024) — XAI	Medical XAI	CXR + CT	90% pneumonia Acc.	Grad-CAM + LIME

3. METHODOLOGY: FASHION-MNIST CNN BASELINE

3.1 Dataset and Preprocessing

Fashion-MNIST [42] is a dataset with 70,000 28x28 grey-scale images, with equal representations of 10 mutually exclusive clothing categories. The official 60,000/ 10,000 train-test partition is adopted. The 60,000 training images are again split into 54,000 real training samples and 6,000 validation samples with stratified random splitting (random state=42) to maintain the same balance of classes in each subset. The pixel intensities are scaled to float32 [0, 1] / 255.0 to ensure gradient values across the backpropagation remain the same [13]. An additional channel dimension is inserted to transform image tensors of (28, 28) into (28, 28, 1), which meets the input requirement of Conv2D. Class labels are represented in one-hot format as 10-dimensional binary vectors in order to be compatible with categorical cross-entropy loss. Fashion-MNIST was selected as the experimental platform consciously. In addition to its usability as a general benchmark, its structural characteristics, such as fixed-format single-channel images, multi-class discrimination with 10 classes, intra-class variability and inter-class visual similarity, have a close relationship with the technical characteristics of real-world image classification tasks such as medical histopathology slide classification, retail product recognition, and satellite image categorisation. The engineering choices tested on the Fashion-MNIST will be directly applicable to these applied fields.

3.2 CNN Architecture

A sequential CNN is built based on three convolutional blocks of different filter depths and then topped with a fully connected classification head. The architecture represents a hierarchy of spatial features abstractions: Block 1 (Conv2D, 32 filters, 3x3, ReLU; MaxPooling2D 2x2) learns the low-level edges and luminance gradient; Block 2 (Conv2D, 64 filters, 3x3, ReLU; MaxPooling2D 2x2) organizes the low-level feature into mid-level texture and contour patterns; Block 3 (Conv2D, 64 filters, 3x3, ReLU; The classification head is a flat version of the feature maps, which is run through a Dense layer of 128 units (ReLU) to the softmax output layer of 10 units. Parameters to be trained: 65,354 - purposely small size, allowing training on non-GPU hardware without requiring any GPU infrastructure. The activation function of ReLU f(x) is used following each convolutional layer, keeping gradient magnitude positive with positive activations and overcoming the vanishing gradient problem of the sigmoid activations of early deep networks [22]. The 2x2 stride of MaxPooling2D reduces spatial dimensions by half at every block, doubling the effective receptive field (of deeper layers) and offering local translation invariance - an important property of a clothing/object classifier that the diagnostic shape features of different image locations can vary. The output of the softmax generates a normalised probability distribution across the 10 classes, allowing a threshold-based decision maker and confidence meaningfully estimated.

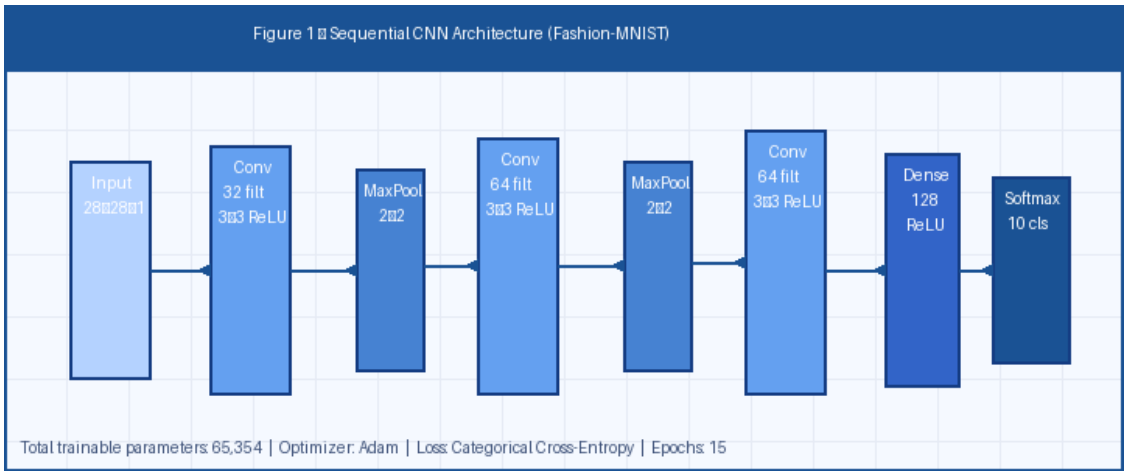

Figure 1. Proposed sequential CNN architecture for Fashion-MNIST multi-class classification. Progressive filter depth (32→64→64) encodes hierarchical spatial feature representations across three convolutional blocks, each followed by 2×2 max-pooling. The classification head maps extracted features through a 128-unit dense layer to a 10-class softmax output. Total parameters: 65,354.

3.3 Training Protocol

The model is compiled with the Adam optimizer [21] at the default learning rate of 0.001, β_1=0.9, β_2=0.999. Adam's adaptive per-parameter moment estimates automatically scale learning rates based on gradient history, enabling rapid initial convergence — evidenced by the sharp validation loss decrease in epochs 1–5 — and robust optimisation through the heterogeneous loss landscape of a 10-class classifier. The loss function is categorical cross-entropy $L=-\Sigma_i\, y_i \log \hat{y}_i$, where y_i is the one-hot true label and $\hat{y}_i$ is the predicted class probability. Training proceeds for 15 epochs with Keras' default batch size of 32. No learning rate scheduling or early stopping is applied in this baseline, making the mild overfitting observed after epoch 10 a deliberate diagnostic finding that motivates the regularisation extensions developed in Section 5.

4. RESULTS AND DISCUSSION

4.1 Training Dynamics

The model is trained using Adam optimizer [21] with the default learning rate of 0.001, b1=0.9, b2=0.999. The adaptive per-parameter moment estimates of Adam have the automatic property of scaling the learning rates by the history of the gradient, allowing learning to converge quickly at the outset, as evidenced by the steep decrease in the validation loss in the first 5 epochs, and the robust optimisation of the non-uniform loss landscape of a 10-class classifier. The loss function is a categorical cross-entropy L [?] S y log y, with y being the one-hot true label and y being the probability of class prediction. The training follows 15 epochs with the default 32 batch size with Keras. This baseline does not use learning rate scheduling or early stopping, and so the mild overfitting that is observed after 10 epochs is a diagnostic feature that is intentional. 4. RESULTS AND DISCUSSION 4.1 Training Dynamics Epoch 15 gives a training accuracy of 94.85 and a validation accuracy of around 90.15 and then levels off, a difference of about 4.7 percentage points that is the main quantitative indicator of the mild overfitting of this shallow

architecture. Validation loss starts with 0.612 at epoch 1 and lowest at epoch 10 with an upward trend until 0.353 at epoch 15 of classic point of validation loss divergence signifying the beginning of memorisation. This behaviour can be explained by two architectural constraints: the convolutional or classification layers of the network did not have dropout regularisation, and the training phase did not have learning rate scheduling to decrease each step size. The two limitations can be directly corrected using standard engineering interventions that are explained in Section 5.

The stable per-epoch gradient variance — standard deviation below 0.02 across all training metrics confirms that the Adam optimizer and batch size of 32 produce a numerically stable training dynamic. This stability is a prerequisite for federated learning deployment: FL requires that each participating client's local training produces convergent gradient estimates that can be meaningfully aggregated, and unstable local training characterised by high loss variance degrades FedAvg aggregation quality.

4.2 Test Set Performance

Testing on the held-out 10,000-sample test set gives a test accuracy of 89.57% and test loss of 0.3562, with a macro-averaged F1-score of 0.90 with all 10 classes. The error rate is 10.43 which translates to 1,043 wrongly classified test samples. These findings align with the literature on benchmarking similar lightweight CNN models [42]; [19] as well as benchmark other small CNNs and form a competitive baseline on which federated and explainability extensions in Sections 5 and 6 are built.

4.3 Per-Class Analysis

The confusion matrix and per-class classification report demonstrate that the performance profile is structured and has direct implications on both the design of federated learning and XAI integration. Categories with morphologically distinct categories have near-perfect F1-scores: the Trouser (F1=0.98), Bag (F1=0.98), and Sandal (F1=0.96) are reliably recognized in all of the sampled test images, and there is only confusion among semantic neighbours within each category. In particular, confusion matrix documents 114 Shirt-T-shirt/top misclassifications and 112 T-shirt/top-Shirt misclassifications - the most common failure mode, which represents about 21 percent of all test errors. Shirt has the lowest F1-score of 0.71 and recall of 0.72, which proves that the representations that the model learned about this classification are not discriminative enough at the 28x28 grayscale. This confusion pattern between classes has two implications. In the case of federated learning: when the training data is distributed to clients with non-identically distributed (non-IID) data partitions (as is systematically true in real-world federated learning, where different institutions are experts on different data subsets) then clients that have a substantial fraction of Shirt-labelled data will buy updates to their gradients that do not converge to the global distribution, poorer federation of a federated model on the globally challenging classes [17]. This is one of the main non-IID data issues in the study of federated learning research (Li et al., 2020). In the case of XAI: it would be motivating to visualise the class activation map [16] of these categories as a Shirt/T-shirt confusion, which would indicate the model is focusing on the right classificative features (collar geometry, hem structure) or spurious (background pixel statistics, image aspect ratio).

Table 2: Per-Class F1-Score with Federated and XAI Research Implications

Class	F1-Score	Confusion Pattern	FL Design Implication	XAI Research Need
Trouser	0.98	Minimal	Well-behaved — stable FL gradient	Grad-CAM: silhouette attention (low priority)
Bag	0.98	Minimal	Reliable anchor class in non-IID partitions	Feature map validates geometric encoding
Sandal	0.96	Minor (Sneaker)	FL convergence aided by distinct features	Grad-CAM: sole/strap region heatmap
Ankle Boot	0.94	Minor (Sneaker)	Stable across client partitions	Attention vs. Sneaker boundary exploration
Dress	0.91	Coat/Pullover	Moderate FL gradient divergence risk	Grad-CAM: hemline vs. collar attention
Pullover	0.88	Coat/Shirt	Non-IID partitions amplify confusion	LIME + Grad-CAM: upper garment discrimination
Coat	0.86	Pullover/Dress	FL non-IID sensitivity — requires FedProx	XAI: collar height attention mapping
T-shirt/top	0.82	Shirt (112 errors)	High FL divergence in non-IID — priority	Grad-CAM: collar/sleeve confusion mapping
Shirt (lowest)	0.71	T-shirt (114 errors)	Critical FL non-IID challenge class	Grad-CAM: misattribution diagnosis (highest priority)

5. FEDERATED LEARNING EXTENSION: PRIVACY-PRESERVING DISTRIBUTED CNN TRAINING

5.1 Federated Learning Architecture

The standard centralised training paradigm that is used in the given study, which is to aggregate all of 54,000 training images on one computational node, and then run a gradient descent algorithm, is not compatible with the data privacy requirements of real-world deployments in healthcare, finance, and legal sectors. Federated Learning [26] addresses this incompatibility by splitting centralised training into a distributed algorithm whereby model weights are distributed and gradient updates are aggregated, whereas raw training data is localised to the institution or device creating the raw training data. The most prevalent FL aggregation algorithm is Federated Averaging (FedAvg), which works in the following manner: (1) a central server initialises global model weights th0 and broadcasts them to K clients participating in the algorithm; (2) each client k runs E epochs of SGD on its local local dataset D k to obtain updated weights th k; (3) the server combines client weights by using a weighted average th t +1 = S k (D k/D k) th k, where |D| = S k D k is the total amount of data per client Within the Fashion-MNIST scenario, it would translate to spreading the 54,000-image training dataset and having it on a series of physically distributed servers (distributed and probably geographically apart) whose gradient aggregation is used instead of centralised batch training [7].

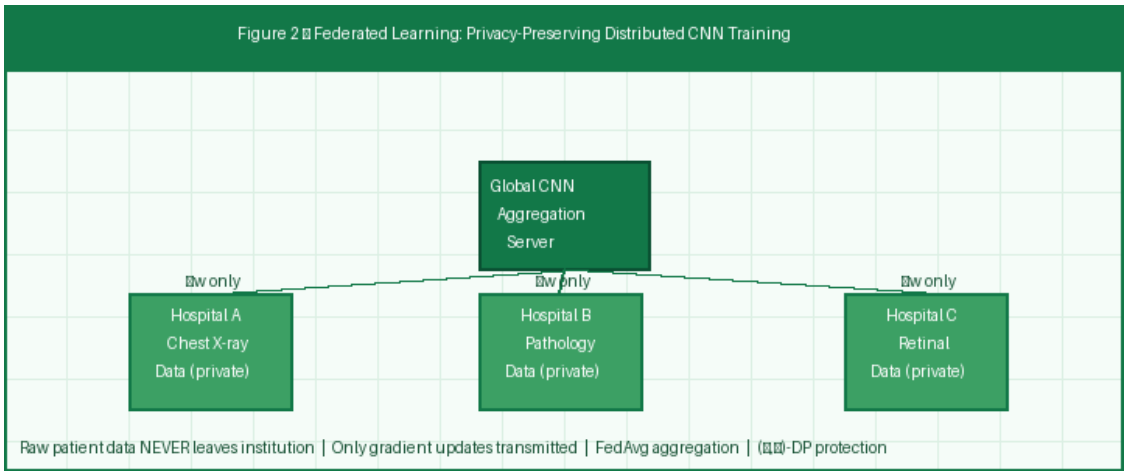

Figure 2. Federated Learning architecture for privacy-preserving distributed CNN training. Each participating institution (Hospital A, Hospital B, Institution C) trains on its private local data partition and transmits only gradient updates — never raw data — to the central aggregation server. FedAvg combines local gradients into a unified global model that achieves performance approaching centralised training while fully preserving data locality.

5.2 Non-IID Data Challenge

Federated CNN training has the main technical difficulty in the non-IID (non-independent and identically distributed) nature of real-world federated partitions. The 54, 000 training images in centralised Fashion-MNIST training are randomly chosen and are of a common balanced distribution of 10 classes. The systematically biased subsets of different clients in a federated environment will be: a footwear retailer will have Sandal, Sneaker and Ankle Boot images; a children clothing store will have T-shirt and Dress images. The resulting heterogeneous local distributions give client gradient updates that are diverting in direction and magnitude, and worsen FedAvg convergence, and give an overall suboptimal model that performs disproportionately on data-rich and poorly underrepresented classes [27]. This non-IID issue is directly manifested in the per-class performance profile of the baseline model. The Shirt/T-shirt misclassification (114 and 112 mutual misclassifications) indicates that even with the centralised training on the balanced Fashion-MNIST dataset, a model that cannot distinguish those two classes is obtained - something that would become far more catastrophic in a federated environment where the clients with Shirt-dominated data form the higher-level gradients that would push the global model towards Shirt consistent decision boundaries, and the client with no Shirt data cannot override it. This is solved by FedProx (Li et al., 2020) with a proximal regularisation term to the local loss of each client that limits the local gradient updates such that they are within a set distance to the global model - decreasing the distance that non-IID data brings to local training. FedNova [40] distributes client gradient updates based on the local amount of training steps, which is then corrected by aggregating the data of clients with different sizes of local datasets. 5.3 Differential privacy Integration. Although FedAvg blocks transmission of raw data, gradient updates of models contain statistical data on the local training data that can potentially support reconstruction attacks - adversarial algorithms that estimate training samples using gradient information [44]. Differential privacy [9] is a formal framework with a mathematically defined notion of privacy [32] that ensures the privacy of individual training examples by adding scaled Gaussian noise to

the gradient updates before the transmission, which limits the information that a single training example can add to the shared model. The DP-FedAvg model is that every client removes local gradients to a constant C L2 limit, followed by the addition of a Gaussian noise N(0, s2C2I), wherein s is the privacy-utility factor. The resultant privacy guarantee is (e, d)-differential privacy - e is the privacy budget (a lower e, the greater the privacy guarantee) and d is the probability of failure. A study done in 2022 [2] using federated DP-CNNs to analyze medical images has shown that clinically acceptable accuracy is achieved at moderate privacy budgets (e[?]10), whereas strict privacy (e[?]1) can also result in diminished accuracy loss by 3-8 percentage points - a trade-off that drives the research directions described in Section 7.

6. XAI INTEGRATION: GRAD-CAM FOR TRANSPARENT CNN DECISION-MAKING

6.1 Motivation: The Black Box Problem

The Fashion-MNIST CNN baseline developed in this paper achieves 89.57% test accuracy through the transformation of 784-dimensional pixel vectors into 10-dimensional probability vectors via a composition of convolutional, pooling, and dense linear operations — a process that, while quantitatively well-characterised, provides no human-interpretable account of why any individual classification decision was made. For the 1,043 misclassified test samples, the model produces incorrect predictions with non-trivial confidence [34]— yet its architecture contains no mechanism for identifying which image regions drove the wrong decision, or which architectural component produced the error. This opacity is unacceptable in regulated deployment contexts: the EU AI Act (2024) mandates explainability for high-risk AI systems, and clinical AI deployment guidelines require that algorithmic decisions be verifiable by qualified human experts.

6.2 Grad-CAM Mathematical Framework

Gradient-weighted Class Activation Mapping (Grad-CAM;[33] generates class-discriminative spatial heatmaps by backpropagating the gradient of the target class score through the network to the final convolutional layer. For a target class c and the final convolutional layer with feature maps $A^k \in R^{(H\times W)}$ (where k indexes filter channels), the Grad-CAM algorithm proceeds as follows:

- Compute the gradient of the class score y^c with respect to each feature map activation: $\partial y^c / \partial A^k_{ij}$
- Compute the importance weight of each feature map by global average pooling of its gradients: $\alpha^c_k = (1/Z) \Sigma_i \Sigma_j (\partial y^c / \partial A^k_{ij})$
- Compute the Grad-CAM heatmap $L^c \in R^{(H\times W)}$ as a weighted combination of feature maps, ReLU-thresholded: $L^c = ReLU(\Sigma_k \alpha^c_k \cdot A^k)$
- Upsample L^c to the original input resolution (28×28) and overlay on the input image to visualise spatially which regions most influenced the classification decision.

The ReLU thresholding in the final step retains only activations with positive influence on the target class — i.e., regions that increase the predicted probability of class c — discarding regions that suppress it. Applied to the Shirt/T-shirt confusion identified in Section 4, Grad-CAM would produce heatmaps revealing whether the model attends to the collar region (a genuine discriminating feature), the overall silhouette (an ambiguous feature shared across upper garments), or peripheral background artifacts (a spurious correlation that should be eliminated through data augmentation or architecture modification) [36]

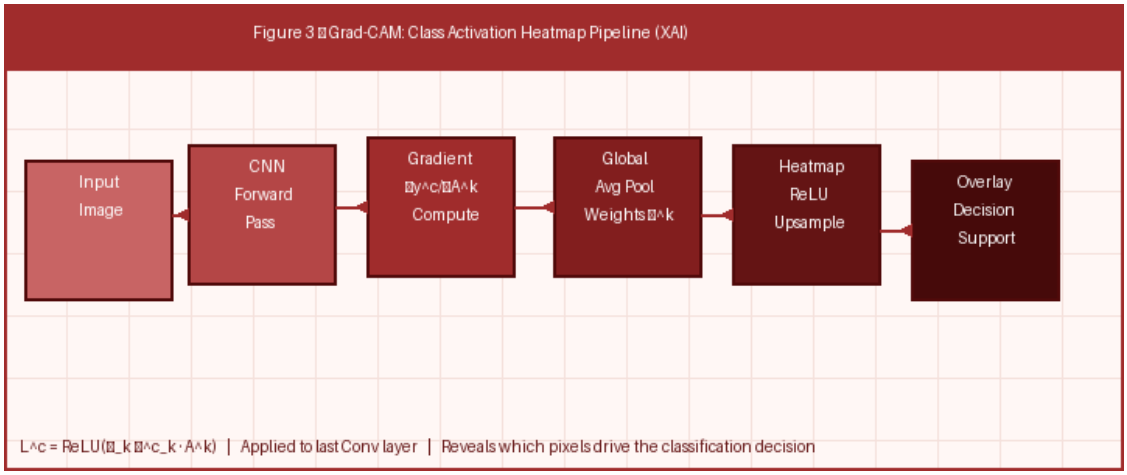

Figure 3. Grad-CAM class activation heatmap pipeline applied to CNN image classification. The gradient of the target class score with respect to final convolutional layer feature maps is computed and spatially averaged to produce importance weights. The resulting heatmap localises the image regions most influential for the model's decision, rendered as a colour-coded overlay on the input image. Applied to Fashion-MNIST, this identifies whether Shirt/T-shirt confusion arises from genuine feature ambiguity or spurious gradient attribution.

6.3 XAI-Augmented Deployment Architecture

Integrating Grad-CAM into the Fashion-MNIST CNN deployment pipeline transforms it from a blackbox accuracy metric into an auditable, interpretable classification system — a prerequisite for any regulated application domain. The XAI-augmented pipeline operates as follows: (1) an input image is passed through the CNN forward pass to generate the predicted class probability vector and intermediate feature maps; (2) the predicted class index is used as the target class c for Grad-CAM computation; (3) gradients are backpropagated to the final convolutional layer (Conv2D with 64 filters) and spatially averaged to produce class-specific importance weights α^c_k; (4) the weighted sum of feature maps is ReLU-thresholded and upsampled to 28×28 to produce the heatmap overlay; (5) the original image, predicted class, confidence score, and Grad-CAM heatmap are jointly presented to a human reviewer for decision validation.

In clinical AI deployment — the most demanding and consequential application context for this architecture — the XAI-augmented pipeline enables radiologists to verify that a CNN pneumonia classification is driven by the expected airspace opacity regions on chest X-ray rather than image

acquisition artifacts or patient demographic proxies. [20] demonstrated in a user study with 26 medical experts that Grad-CAM visualisations on pneumonia and COVID-19 classifiers achieve positive ratings for clinical relevance and comprehensibility — validating the practical utility of this XAI integration beyond its theoretical properties [6].

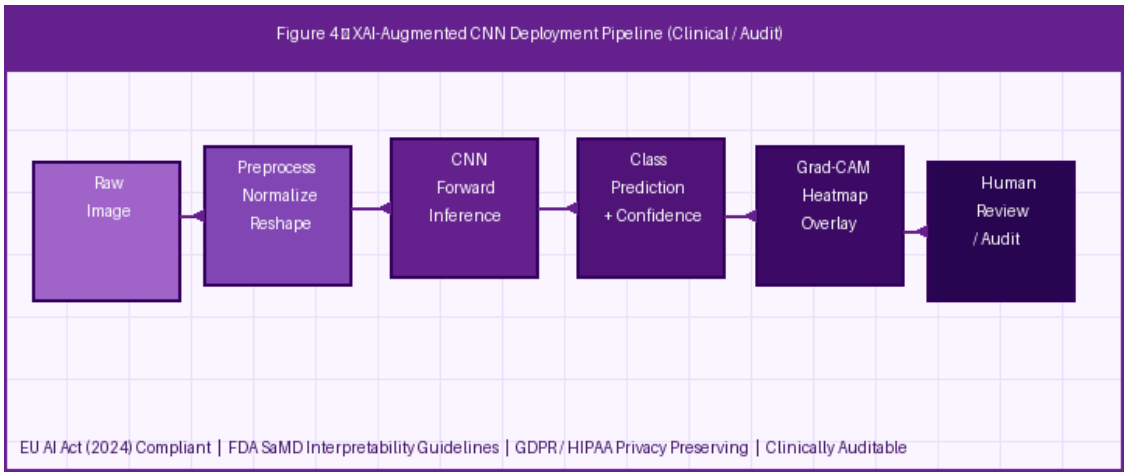

Figure 4. XAI-augmented CNN deployment pipeline for transparent, auditable classification. The raw input image passes through standard preprocessing and CNN forward inference to produce a class prediction and confidence score. Simultaneously, Grad-CAM backpropagation generates a spatial heatmap identifying the image regions most influential for the classification decision. Both outputs are presented jointly to a human reviewer — enabling regulatory compliance, error diagnosis, and clinical validation in high-stakes deployment contexts.

7. FUTURE RESEARCH DIRECTIONS

7.1 Federated Learning with Non-IID Robustness

The added Fashion-MNIST CNN deployment pipeline with Grad-CAM turns it into a blackbox-accurate system despite being made interpretable and auditable classifier - a condition to any regulated domain of application. The XAI-augmented pipeline works in the following way: (1) an input image is fed through the CNN forward pass to obtain the probability of the predicted class and intermediate feature maps; (2) the predicted index of the class is used as the target class c to compute the Grad-CAM; (3) the backpropagated gradients [3] are zeroed to the final convolutional layer (Conv2D with 64 filters) and spatially averaged to obtain the class-specific importance weights a ck; (4) the weighted sum of the feature maps is ReLU-thresholded

In clinical AI adoption, the application area of AI that demands the greatest amount of application effort and has the most consequential outcomes, the XAI-enhanced pipeline will allow radiologists to ensure that a CNN pneumonia classifier is trained to be motivated by the desired airspace opacities on chest X-ray and not by image acquisition artifacts or patient demographic surrogates. In a user study involving 26 medical experts, [20] proved that Grad-CAM visualisations of pneumonia and COVID-19 classifiers are

rated positively on clinical relevance and comprehensibility - which confirms the theoretical proposals of this XAI integration as useful in practice.

7.1 Federated Learning Non-IID Robustness

The most readily productive extension of the research is the introduction and testing of the non-IID-robust versions of federated learning on a partitioned Fashion-MNIST benchmark. FedProx (Li et al., 2020) introduces a proximal regularisation term m/2 ||w[?]w^t||2 to the local client loss, and the gradient updates are limited to stay within a localised neighbourhood of the global model - greatly enhancing convergence when there is heterogenous data distribution. To address client drift, a systematic bias that non-IID local training causes the global model, SCAFFOLD [19] proposes control variates as a means of correcting it. Strict comparative analysis of FedAvg, FedProx, FedNova and SCAFFOLD on Fashion-MNIST with systematically tuned non-IID partitioning strategies - ranging in heterogeneity (mildly) between Dirichlet a=0.5 to severely heterogeneous (each client has data in only 2 of 10 classes) - would give engineering actionable advice on federated deployment of CNN classifiers in non-experimental data distributions.

7.2 Optimisation of Privacy-Utility Trade-Off

The addition of differential privacy to federated CNN training provides a basic trade-off between the strength of privacy (controlled by the noise multiplier s and privacy budget e) and model utility (classification accuracy). Empirical findings up to date suggest that moderate privacy budgets (e[?]10) maintain clinical accuracy and strict budgets (e[?]1) lower accuracy by 3-8 percentage points [2]; Marnissi et al., 2025. Future studies should examine adaptive noise scheduling schemes where s decreases with training and the model is approaching convergence and that ensures e-DP guarantees and minimises the cost of noise injection. Additional privacy amplification techniques, such as subsampling [4] and secure aggregation [5] lower the privacy cost per training round, and need to be compared to adaptive DP in Fashion-MNIST and medical imaging benchmarks.

7.3 Advanced XAI: Beyond Grad-CAM

Although Grad-CAM offers the benefit of offering spatial attribution of the CNN-based classifier, its gradient based mechanism offers heatmaps which are averaged across all samples of the input - constraining its use in understanding the individual misclassifications. This limitation is overcome by a number of developed XAI techniques. There is proposed SHAP (SHapley Additive exPlanations; [25] which is a theoretically motivated feature attribution that has guaranteed local accuracy, missingness and consistency properties unmet by Grad-CAM. The LIME (Local Interpretable Model-Agnostic Explanations; [30] produces instance-specific attributions through the use of a local linear surrogate model, which is trained on perturbed input samples; it offers pixel-level attribution to individual misclassified images. Grad-CAM does not have axiomatic completeness, which is guaranteed with Integrated Gradients [38]. Applying a complete XAI benchmark on the cluster of Shirt/T-shirt misclassification errors, between Grad-CAM, Grad-CAM++, SHAP, LIME, and Integrated Gradients to

understand the 226 mutual Shirt/T-shirt errors, would reveal the most diagnostically useful attribution weight to CNN error analysis and architectural tuning.

7.4 Vision Transformer Application

Vision Transformers ViT; [8] - a model that treats image patches as sequential tokens, where the input undergoes multi-head self-attention, outperform CNNs on large-scale benchmarks but fail on small ones, such as Fashion-MNIST (28x28, 70,000 images), because they do not have convolutional inductive biases to extract local spatial features. Hybrid CNN-ViT systems (such as CvT, EfficientFormer) in which initial convolutional stages encode local features, followed by subsequent global feature encoding by transformer attention layers, represent a good direction towards higher accuracy than pure CNNs on Fashion-MNIST without sacrificing the interpretability of the architectural encoding/decoding process that pure CNNs enjoy with Grad-CAM. The self-attention approach of ViT has also offered a natural architecture-based explainability approach, attention rollout maps, in federated environments, that serves as a complement to Grad-CAM to provide post-hoc visual attribution.

7.5 Cross-Domain Transfer to Medical Imaging Benchmarks

The final test of the federated XAI-CNN model that has been constructed in the present paper is going to be its application to the actual medical imaging standards. MedMNIST [43] offers a standardised set of 18 medical image tasks in the same format as Fashion-MNIST (28x28) with classification tasks related to dermoscopy, X-ray, OCT, ultrasound, and pathology, so that they can be directly transferred to architectural models without change. Implementing federated CNN baseline training on both centralised and federated settings of PathMNIST (blood cell pathology, 9 classes) and ChestMNIST (chest X-ray, 14 classes) on the DP and Grad-CAM XAI would offer the first overall assessment of the entire privacy-preserving, explainable CNN pipeline on standardised medical benchmarking, leading to the generation of the empirical validation necessary to consider such pipelines in clinical settings.

8. CONCLUSION

The three-layer discussion in this paper provided the foundational engineering capabilities by defining the challenge of image classification, the extension framework to Federated Learning overcoming the privacy obstacles to practical implementation, and the architecture of a Grad-CAM XAI integration to fulfill the interpretability demands of controlled application environments of high stakes.The Fashion-MNIST baseline — a sequential CNN with 3 convolutional blocks (32/64/64 filters), 65,354 parameters, trained for 15 epochs with Adam optimisation and categorical cross-entropy — achieves 89.57% test accuracy and a macro F1-score of 0.90. Per-class analysis identifies the Shirt/T-shirt confusion cluster (114 and 112 mutual misclassifications; Shirt F1=0.71) as the primary failure mode, attributable to the fundamental information-theoretic limits of 28×28 grayscale imagery for fine-grained garment discrimination. The mild overfitting observed after epoch 10 — a 4.7% train-validation accuracy gap — is directly addressable through dropout regularisation, early stopping, and data augmentation.

The Federated Learning framework maps these findings onto a distributed training architecture in which FedAvg aggregates gradients from geographically decentralised clients without exposing raw training

data, with differential privacy providing formal (ε, δ)-DP guarantees against gradient reconstruction attacks. The non-IID data challenge — amplified by the same inter-class visual similarity that produces the Shirt/T-shirt confusion in centralised training — is identified as the critical engineering bottleneck for federated CNN deployment, motivating FedProx and SCAFFOLD as immediate research extensions.

The Grad-CAM XAI integration transforms the baseline from a blackbox classifier into an auditable, spatially interpretable decision system that meets the regulatory explainability requirements of the EU AI Act (2024) and clinical AI deployment guidelines. Together, these extensions define a technically rigorous, research-worthy, and socially consequential agenda: the development of federated, privacy-compliant, and human-interpretable CNN classifiers for deployment in healthcare, finance, and other domains where data sensitivity and decision accountability are non-negotiable requirements.

REFERENCES

[1] Abu, M. A., Indra, N. H., Rahman, A. H. A., Sapiee, N. A., & Ahmad, I. (2019). A study on image classification based on deep learning and TensorFlow. International Journal of Engineering Research and Technology, 12(4), 563–569.

[2] Adnan, M., Kalra, S., Cresswell, J. C., Taylor, G. W., & Tizhoosh, H. R. (2022). Federated learning and differential privacy for medical image analysis. Scientific Reports, 12, 1953. https://doi.org/10.1038/s41598-022-05934-2

[3] Almazroi, A. A., Alsubaei, F. S., Ayub, N., & Jhanjhi, N. Z. (2024). Inclusive Smart Cities: IoT-Cloud Solutions for Enhanced Energy Analytics and Safety. *International Journal of Advanced Computer Science & Applications*, *15*(5).

[4] Balle, B., Barthe, G., Gaboardi, M., Hsu, J., & Sato, T. (2020). Hypothesis testing interpretations and renyi differential privacy. Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics (AISTATS), PMLR 108, 2496–2506.

[5] Bonawitz, K., Ivanov, V., Kreuter, B., Marcedone, A., McMahan, H. B., Patel, S., ... & Seth, K. (2017). Practical secure aggregation for privacy-preserving machine learning. Proceedings of the 2017 ACM SIGSAC Conference on CCS, 1175–1191.

[6] Brohi, S. N., Jhanjhi, N. Z., Brohi, N. N., & Brohi, M. N. (2020). Key Applications of State-of-the-Art Technologies to Mitigate and Eliminate COVID-19.pdf.. https://doi.org/10.36227/techrxiv.12115596.v1

[7] Chandra Mallojjala, S., Sarkar, R., Karugu, R. W., Manna, M. S., Ray, S., Mukherjee, S., & Hirschi, J. S. (2022). Mechanism and origin of remote stereocontrol in the organocatalytic enantioselective formal c (sp2)–h alkylation using nitroalkanes as alkylating agents. *Journal of the American Chemical Society*, *144*(38), 17399-17406.

[8] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., ... & Houlsby, N. (2021). An image is worth 16×16 words: Transformers for image recognition at scale. International Conference on Learning Representations (ICLR).

[9] Dwork, C., Roth, A., et al. (2014). The algorithmic foundations of differential privacy. Foundations and Trends in Theoretical Computer Science, 9(3–4), 211–407.

[10] European Union. (2016). General Data Protection Regulation (GDPR). Official Journal of the European Union, L119, 1–88.

[11] European Union. (2024). Artificial Intelligence Act. Official Journal of the European Union.

[12] FDA. (2021). Artificial intelligence/machine learning (AI/ML)-based software as a medical device (SaMD) action plan. U.S. Food and Drug Administration.

[13] Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep learning. MIT Press.

[14] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. Proceedings of IEEE CVPR, 770–778.

[15] IBM Security. (2023). Cost of a data breach report 2023. IBM Corporation.

[16] JingXuan, C., Tayyab, M., Muzammal, S. M., Jhanjhi, N. Z., Ray, S. K., & Ashfaq, F. (2024, November). Integrating AI with robotic process automation (RPA): advancing intelligent automation systems. In *2024 IEEE 29th Asia Pacific Conference on Communications (APCC)* (pp. 259-265). IEEE.

[17] Khalil, M.I., Humayun, M., Jhanjhi, N.Z., Talib, M.N., Tabbakh, T.A. (2021). Multi-class Segmentation of Organ at Risk from Abdominal CT Images: A Deep Learning Approach. In: Peng, SL., Hsieh, SY., Gopalakrishnan, S., Duraisamy, B. (eds) Intelligent Computing and Innovation on Data Science. Lecture Notes in Networks and Systems, vol 248. Springer, Singapore. https://doi.org/10.1007/978-981-16-3153-5_45

[18] Kadam, S. S., & Adamuthe, A. C. (2021). CNN model for image classification on MNIST and Fashion-MNIST dataset. Journal of Emerging Technologies and Innovative Research, 8(5).

[19] Karimireddy, S. P., Kale, S., Mohri, M., Reddi, S. J., Stich, S. U., & Suresh, A. T. (2020). SCAFFOLD: Stochastic controlled averaging for federated learning. Proceedings of ICML 2020.

[20] Kassem, A. A., et al. (2024). Evaluating explainable artificial intelligence (XAI) techniques in chest radiology imaging through a human-centered lens. PMC / PLOS ONE.

[21] Kingma, D. P., & Ba, J. (2015). Adam: A method for stochastic optimization. International Conference on Learning Representations (ICLR). https://arxiv.org/abs/1412.6980

[22] Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. Advances in Neural Information Processing Systems, 25, 1097–1105.

[23] LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. Proceedings of the IEEE, 86(11), 2278–2324.

[24] Li, T., Sahu, A. K., Zaheer, M., Sanjabi, M., Smola, A., & Smith, V. (2020). Federated optimization in heterogeneous networks (FedProx). Proceedings of MLSys 2020.

[25] Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions (SHAP). Advances in Neural Information Processing Systems, 30.

[26] McMahan, H. B., Moore, E., Ramage, D., Hampson, S., & y Arcas, B. A. (2017). Communication-efficient learning of deep networks from decentralized data. Proceedings of AISTATS 2017.

[27] Ninama, H., Raikwal, J., Ravuri, A., Sukheja, D., Bhoi, S. K., Jhanjhi, N. Z., ... & Abdelmaboud, A. (2024). Computer vision and deep transfer learning for automatic gauge reading detection. *Scientific Reports*, *14*(1), 23019.

[28] Nocentini, O., Kim, J., Bashir, M. Z., & Cavallo, F. (2022). Image classification using multiple convolutional neural networks on the Fashion-MNIST dataset. Sensors, 22(23), 9544.

[29] Rajpurkar, P., Irvin, J., Zhu, K., Yang, B., Mehta, H., Duan, T., ... & Ng, A. Y. (2017). CheXNet: Radiologist-level pneumonia detection on chest X-rays with deep learning. arXiv:1711.05225.

[30] Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). 'Why should I trust you?': Explaining the predictions of any classifier (LIME). Proceedings of KDD 2016, 1135–1144.

[31] Rieke, N., Hancox, J., Li, W., Milletari, F., Roth, H. R., Albarqouni, S., ... & Cardoso, M. J. (2020). The future of digital health with federated learning. npj Digital Medicine, 3(1), 119.

[32] Saeed, S., Jhanjhi, N. Z., Khan, M. A., & Yadav, D. K. (2025). Digital transformation and cybersecurity challenges. *Frontiers in Computer Science*, *7*, 1631362.

[33] Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-CAM: Visual explanations from deep networks via gradient-based localization. Proceedings of IEEE ICCV, 618–626.

[34] Shah, I. A., Jhanjhi, N. Z., & Ray, S. K. (2024). IoT devices in drones: security issues and future challenges. In *Cybersecurity Issues and Challenges in the Drone Industry* (pp. 217-235). IGI Global Scientific Publishing.

[35] Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556.

[36] S. M. Muzammal, R. K. Murugesan, N. Z. Jhanjhi and L. T. Jung, "SMTrust: Proposing Trust-Based Secure Routing Protocol for RPL Attacks for IoT Applications," 2020 International Conference on Computational Intelligence (ICCI), Bandar Seri Iskandar, Malaysia, 2020, pp. 305-310, doi: 10.1109/ICCI51257.2020.9247818.

[37] Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: A simple way to prevent neural networks from overfitting. Journal of Machine Learning Research, 15(1), 1929–1958.

[38] Sundararajan, M., Taly, A., & Yan, Q. (2017). Axiomatic attribution for deep networks. Proceedings of ICML 2017.

[39] Teo, Z. L., et al. (2024). Federated machine learning in healthcare: A systematic review on clinical applications and technical architecture. Cell Reports Medicine, 5.

[40] Wang, J., Liu, Q., Liang, H., Joshi, G., & Poor, H. V. (2020). Tackling the objective inconsistency problem in heterogeneous federated optimization (FedNova). Advances in NeurIPS 33.

[41] Williamson, S. M., & Prybutok, V. (2025). Privacy-preserving federated learning for collaborative medical data mining in multi-institutional settings. Scientific Reports. https://doi.org/10.1038/s41598-025-97565-4

[42] Xiao, H., Rasul, K., & Vollgraf, R. (2017). Fashion-MNIST: A novel image dataset for benchmarking machine learning algorithms. arXiv:1708.07747.

[43] Yang, J., Shi, R., Wei, D., Liu, Z., Zhao, L., Ke, B., ... & Ni, B. (2023). MedMNIST v2: A large-scale lightweight benchmark for 2D and 3D biomedical image classification. Scientific Data, 10, 41.

[44] Zhu, L., Liu, Z., & Han, S. (2019). Deep leakage from gradients. Advances in Neural Information Processing Systems, 32.