Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

# FusionNet Lite: Compact Federated Deep Learning for Predictive Maintenance under Label Heterogeneity and Temporal Generalization

**Aman Sharma**

104893880@student.swin.edu.au

Swinburne University of Technology

**Kwan Yong Sim**

Swinburne University of Technology

**Sivachandran Chandrasekaran**

Swinburne University of Technology

**Research Article**

**Keywords:**

**Posted Date:** November 9th, 2026

**DOI:** https://doi.org/10.21203/rs.3.rs-10856723/v1

**License:** This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

**Additional Declarations:** No competing interests reported.

# FusionNet Lite: Compact Federated Deep Learning for Predictive Maintenance under Label Heterogeneity and Temporal Generalization

Aman Sharma[1*], Kwan Yong Sim[2] and Sivachandran Chandrasekaran[1]

[1*]School of Science, Computing and Emerging Technologies, Swinburne University of Technology, Hawthorn, Australia.
[2]Faculty of Engineering, Computing and Science, Swinburne University of Technology, Sarawak Campus, Malaysia.

*Corresponding author(s). E-mail(s): 104893880@student.swin.edu.au;
Contributing authors: ksim@swinburne.edu.my;
schandrasekaran@swin.edu.au;

**Abstract**

Federated learning (FL) supports collaborative predictive maintenance (PdM), but compactness alone does not establish reliable failure detection or consistent explanations. We evaluate FusionNet Lite, a 1,420-parameter feature-axis classifier, against four lightweight baselines under centralized training and Federated Averaging (FedAvg). Ten matched core seeds and validation-selected checkpoints are used with stratified AI4I 2020 partitions and a fixed-row chronological MetroPT3 holdout. Separate studies extend the three-client core to ten clients and examine architectural, optimization and participation sensitivities. On MetroPT3, FedAvg FusionNet Lite achieves area under the receiver operating characteristic curve (AUROC) **0.9351 ± 0.0462** and area under the precision–recall curve (AUPRC) **0.2197±0.1971**, expressed as mean and sample standard deviation. Switching from original-grid validation-$F_1$ to expanded-grid validation-$F_2$ selection changes mean test $F_2$ from **0.0423 ± 0.0995** to **0.1566 ± 0.2044**, retaining substantial variability. FusionNet Lite has the smallest parameter count and lowest fixed-batch parameter–latency index among the five core architectures, while predictive rankings depend on dataset and configuration. Paired, class-specific Integrated Gradients (IG) comparisons across all five models show variable central–federated agreement. These findings support

a conditional performance–complexity trade-off within a controlled benchmark. Forecasting lead time, physical energy savings and industrial deployment readiness remain unestablished; future work requires independent equipment, alternative aggregation and device-level validation.

**Keywords:** Federated Learning, Predictive Maintenance, Lightweight Deep Learning, Non-IID Data, Temporal Generalization, Integrated Gradients

# 1 Introduction

Predictive maintenance (PdM) uses operational data to support equipment monitoring and failure detection, but industrial measurements are often distributed across machines and organizations. Federated learning (FL) enables collaborative training while retaining raw observations locally, making it relevant to manufacturing and maintenance applications [1, 2]. Current approaches pursue communication-efficient learning [3], resource-conscious edge designs [4, 5] and physical-device evaluation [6]. Nevertheless, non-independent and non-identically distributed (non-IID) client data complicate optimization [7, 8], while temporal dependence and rare failures affect evaluation. Explainability studies in PdM and federated monitoring add another requirement: predictions should also be examined through feature attributions [9, 10]. The practical question is therefore whether a compact model can retain useful detection, measured computational efficiency and explanatory agreement under explicitly defined data and training conditions.

FusionNet Lite combines separable convolution, ConvMixer-style feature mixing, residual connections and squeeze-and-excitation (SE) channel reweighting in a 1,420-parameter classifier operating on individual observations [11, 12]. We compare it with convolutional neural network (CNN), bidirectional long short-term memory (BiLSTM), multilayer perceptron (MLP) and CNN–long short-term memory (CNN–LSTM) baselines under centralized training and Federated Averaging (FedAvg), using ten matched core seeds and comparable validation-based checkpoint selection. AI4I 2020 uses stratified partitions, whereas MetroPT3 uses a fixed-row chronological 55/20/25 split. Paired, class-specific Integrated Gradients (IG) comparisons assess central–federated attribution agreement across all five architectures.

The study examines whether compact parameterization can retain useful detection and explanation agreement under federation. Its contributions combine architectural design with controlled empirical evaluation:

- A compact 1,420-parameter architecture using established convolution, feature-mixing, residual and SE operations, assessed through component ablations and a standard-convolution comparison.
- Matched ten-seed core comparisons across five architectures and two datasets, with validation-selected checkpoints and separate analyses of decision thresholds and imbalance-sensitive losses.
- Sensitivity analyses of feature order, learning rate, label allocation, client count and participation under declared experimental conditions.

- Class-specific, paired IG comparisons across all five architectures, with resource measurements that distinguish explanatory agreement, detection quality, and computational cost.

Section 2 contrasts these contributions with related studies and identifies the evaluation gaps. Section 3 then defines the data partitions, model implementations and validation-only selection rules used to examine them.

## 2 Literature Review

Existing FL research addresses complementary parts of this evaluation problem. Surveys examine decentralized topologies and statistical heterogeneity [13, 14], while communication-efficient learning and physical-device studies address model exchange and resource constraints [3, 6]. With full-model updates at fixed numerical precision, fewer parameters reduce the nominal payload per exchange; actual communication cost also depends on protocol overhead and participation. Client partitioning introduces a separate distinction. FedLabSync uses label synchronization and feed-forward networks for hybrid sample and feature partitions in Bosch product inspection and aircraft remaining useful life (RUL) estimation using Commercial Modular Aero-Propulsion System Simulation (C-MAPSS) data [15]. That setting differs from the shared-feature, label-Dirichlet FedAvg comparison here. These studies support examining realized client distributions and measured resource use, without assuming that architectural compactness resolves every form of heterogeneity.

Representation and evaluation further limit comparisons between studies. Davari et al. investigated sparse-autoencoder alarms for a train air production unit (APU) [16]; Veloso et al. described a separate 2022 MetroPT recording and interval-based failure evaluation [17]. That recording does not establish the annotation provenance of the 2020 MetroPT-3 input used here [18]. The published FusionNet framework combines CNN, BiLSTM and MLP branches over sensor windows, with random splitting of overlapping MetroPT3 windows [19]. Its results are consequently not directly comparable with the present single-observation chronological evaluation and simulated federation. The early FusionNet Lite preprint documents an earlier version of this same study [20]. Related explainability work in PdM and federated intrusion detection provides context for attribution analysis [9, 10, 21]; here, we ask whether class-paired attribution agreement accompanies useful detection across matched architectures. Table 1 summarizes these differences in data, models, federation and evaluation.

These distinctions lead to the shared preprocessing, validation-based selection and matched model comparisons specified in Sect. 3.

**Table 1** Scope of related studies and the present evaluation.

| Study | Data and model | Federation and partitioning | Evaluation focus |
|---|---|---|---|
| Davari et al. [16] | Train APU sensors; sparse and variational autoencoders. | Reconstruction-based anomaly detection rather than a federated-client comparison. | Filtered sensor alarms; differs from supervised row-level classification. |
| Veloso et al. [17] | MetroPT 2022 compressor recording; dataset description. | Dataset description and maintenance-interval definitions; no shared centralized–FedAvg benchmark. | Interval-based failure detection and warning requirements; distinct from the 2020 input used here. |
| FedLabSync [15] | Bosch inspection and C-MAPSS RUL; feed-forward networks. | Hybrid sample/feature partitions; label synchronization and prediction exchange. | Classification and RUL across partition settings; differs from shared-feature FedAvg. |
| Wong et al. [6] | Learning testbed on physical edge devices. | Federated training with device and data heterogeneity. | Learning, computation and communication costs; complements simulated-client evaluations. |
| FusionNet [19] | MetroPT3, AI4I and C-MAPSS; CNN–BiLSTM–MLP sequence fusion. | Sequence-model evaluation with randomly split overlapping MetroPT3 windows. | Predictive sequence modelling; differs in representation and temporal partitioning. |
| Present study | AI4I and MetroPT3; compact feature-axis classifier and four baselines. | Centralized and synchronous FedAvg; three-client core, ten-client and participation sensitivities. | Ten-seed core, matched checkpoints, validation-selected thresholds, resource measurements and class-paired IG. |

Note: The table compares study scope; differing protocols preclude a direct numerical ranking.

# 3 Methodology

The experimental framework compares FusionNet Lite with four reference architectures under centralized training and simulated-client FedAvg. Following dataset-specific preprocessing, the core comparisons use ten matched fresh seeds, with the same validation-based checkpoint criterion applied in both optimization modes. Separate controlled studies examine architectural components, feature ordering, client allocation, loss function, learning rate, and partial participation. The evaluation connects predictive discrimination and operating-point performance with measured inference latency and paired central–federated attribution diagnostics. Fig. 1 summarizes these stages; we retain the archived five-seed campaign as reference evidence and do not pool it with the new experiments.

## 3.1 FusionNet Lite architecture

*FusionNet Lite* is a compact classifier that processes one preprocessed observation at a time. The input has shape $(F, 1)$, with the sequence axis indexing retained features, not successive timestamps. Convolution mixes adjacent features in a specified order. Chronological train–test evaluation assesses performance on later observations but does not make the model a temporal forecaster. Both optimization modes use the same full-model builder.

The feature extractor applies two SeparableConv1D layers with 16 filters and kernel size three. Each convolution includes a ReLU activation and is followed by Layer

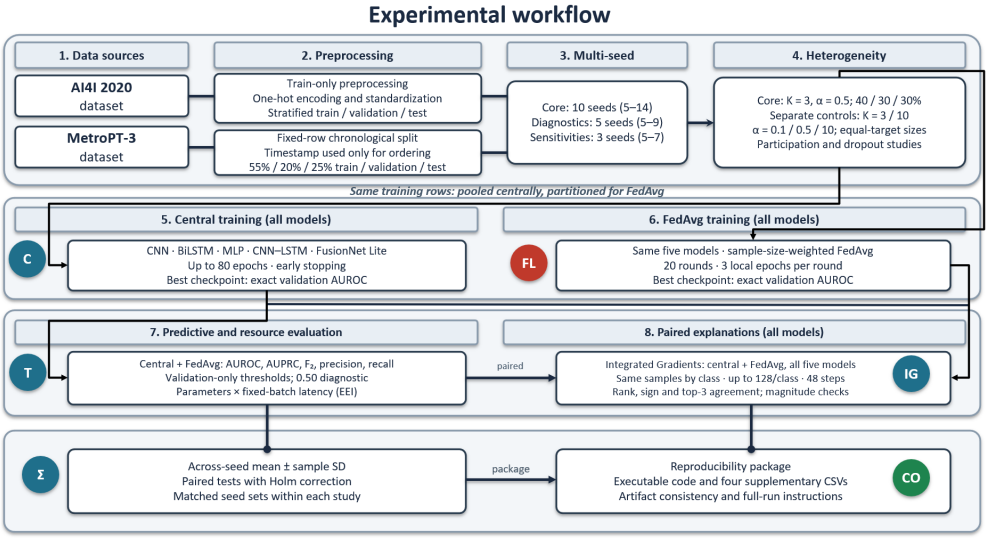



**Fig. 1** Overall methodological workflow of the revised study.

Normalization. Before residual addition, a linear $1\times1$ Conv1D projects the one-channel input to 16 channels. The resulting representation enters a ConvMixer-style block with a ReLU-activated depthwise convolution, Layer Normalization, a ReLU-activated pointwise convolution, and a second Layer Normalization. A second residual addition precedes SE channel reweighting and the compact classification head shown in Fig. 2.

The feature-mixing and channel-reweighting components adapt established ConvMixer and SE ideas [11, 12]; the contribution lies in their compact composition and controlled evaluation, not in a new primitive operator.

Channel refinement uses a squeeze-and-excitation (SE) mechanism. Global average pooling compresses the mixed representation into a channel descriptor, which $\mathrm{Dense}(8, \mathrm{ReLU})$ and $\mathrm{Dense}(16, \mathrm{sigmoid})$ transformations process. The resulting coefficients reweight the feature channels before classification. The final prediction head applies global average pooling, $\mathrm{Dropout}(0.1)$, $\mathrm{Dense}(16, \mathrm{ReLU})$, and a single sigmoid output unit for binary failure prediction.

The evaluated FusionNet Lite configuration contains 1,420 trainable parameters. Its compactness is particularly relevant to FedAvg because clients and the aggregation server repeatedly exchange full model weights. The architectural objective is therefore not to assume predictive superiority from parameter reduction, but to determine whether a substantially smaller model can provide a useful predictive-performance and efficiency trade-off when assessed under the same protocol as larger lightweight baselines.

Algorithm 1 summarises the principal architectural operations. The pseudocode provides a framework-independent representation of the forward path, while the executable implementation is included in the reproducibility package.

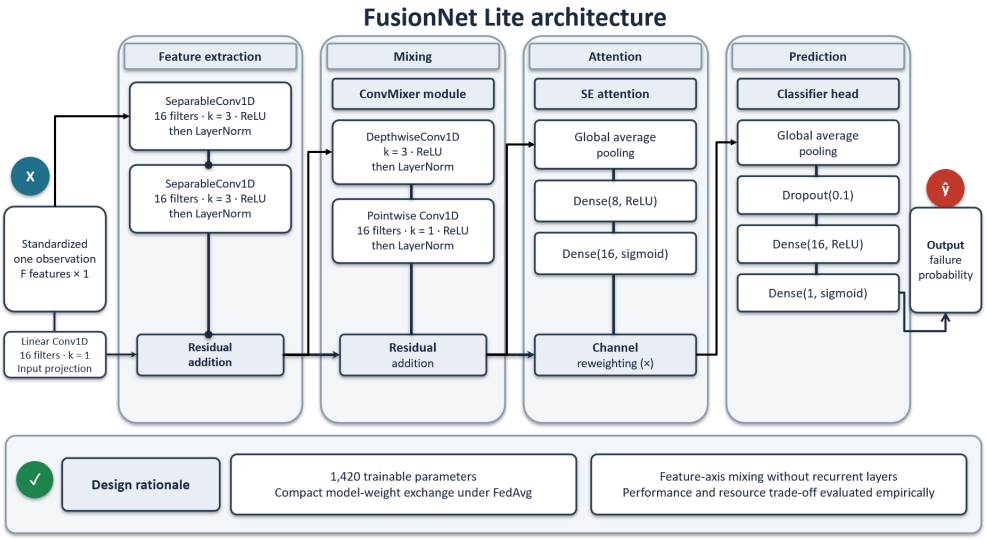



**Fig. 2** FusionNet Lite architecture.

**Table 2** Layer-wise configuration of FusionNet Lite.

| Stage | Configuration |
|---|---|
| Feature extraction 1 | SeparableConv1D (16 filters, $k$=3, ReLU), then LayerNorm |
| Feature extraction 2 | SeparableConv1D (16 filters, $k$=3, ReLU), then LayerNorm |
| Residual pathway | Linear Conv1D (16 filters, $k$=1, same padding) projects the input before addition to the second feature-extraction output |
| ConvMixer module | DepthwiseConv1D ($k$=3, ReLU), then LayerNorm; pointwise Conv1D (16 filters, $k$=1, ReLU), then LayerNorm; residual addition |
| SE attention | GlobalAveragePooling1D → Dense(8, ReLU) → Dense(16, Sigmoid); channel-wise scaling |
| Classification head | GlobalAveragePooling1D → Dropout(0.1) → Dense(16, ReLU) → Dense(1, Sigmoid) |
| Trainable parameters | 1,420 |

**Algorithm 1** FusionNet Lite forward computation

**Require:** Standardised input tensor $x$ of shape $(F, 1)$

**Feature extraction**
$h_1 \leftarrow \text{SeparableConv1D}(16, k{=}3, \text{padding} = \text{"same"}, \text{activation} = \text{ReLU})(x)$
$h_1 \leftarrow \text{LayerNorm}(h_1)$
$h_2 \leftarrow \text{SeparableConv1D}(16, k{=}3, \text{padding} = \text{"same"}, \text{activation} = \text{ReLU})(h_1)$
$h_2 \leftarrow \text{LayerNorm}(h_2)$

**Residual feature propagation**
$q \leftarrow \text{Conv1D}(16, k{=}1, \text{padding} = \text{same}, \text{activation} = \text{linear})(x)$
$r \leftarrow h_2 + q$

**ConvMixer block**
$c \leftarrow \text{DepthwiseConv1D}(3, \text{padding} = \text{"same"}, \text{activation} = \text{ReLU})(r)$
$c \leftarrow \text{LayerNorm}(c)$
$p \leftarrow \text{Conv1D}(16, k{=}1, \text{activation} = \text{ReLU})(c)$
$p \leftarrow \text{LayerNorm}(p)$
$m \leftarrow r + p$

**SE attention**
$s \leftarrow \text{GlobalAveragePooling1D}(m)$
$s \leftarrow \text{Dense}(8, \text{activation} = \text{ReLU})(s)$
$s \leftarrow \text{Dense}(16, \text{activation} = \text{sigmoid})(s)$
$s \leftarrow \text{Reshape}((1, 16))(s)$
$a \leftarrow m \odot s$

**Classification head**
$z \leftarrow \text{GlobalAveragePooling1D}(a)$
$z \leftarrow \text{Dropout}(0.1)(z)$
$z \leftarrow \text{Dense}(16, \text{activation} = \text{ReLU})(z)$
$\hat{y} \leftarrow \text{Dense}(1, \text{activation} = \text{sigmoid})(z)$
**return** $\hat{y}$

## 3.2 Baseline architectures

Four reference architectures separate the proposed composition from simpler convolutional, recurrent and fully connected alternatives. The CNN contains one Conv1D layer with 16 filters, kernel size three, same padding and ReLU activation, followed by max pooling of size two, flattening, Dense(32, ReLU) and a sigmoid output. The BiLSTM applies a bidirectional LSTM with 16 units in each direction, followed by Dense(32, ReLU) and the binary output. CNN–LSTM uses Conv1D(16, 3, ReLU), max pooling of size two with same padding, LSTM(16) and Dense(32, ReLU) before classification. The MLP uses hidden layers of 64, 32 and 16 ReLU units. All models end in a single sigmoid unit.

For convolutional and recurrent models, the sequence axis traverses the ordered features of one observation; the MLP receives the same features as a flat vector. These models serve as compact architectural references and are not matched by parameter count. Input-dependent parameter counts are reported alongside performance and timing. Shared optimization settings define a controlled benchmark but do not establish that every architecture is optimally tuned; the bounded learning-rate study addresses that separate question.

## 3.3 Datasets and leakage-aware preprocessing

Two PdM datasets are used: AI4I 2020 and MetroPT3. Although both are transformed into the same model-input format, their partitioning strategies differ because their underlying data-generating structures are different.

For AI4I 2020, the binary target is the logical OR of TWF, HDF, PWF, OSF and RNF. We remove these indicators and the original *Machine failure* column before modelling, along with UDI and Product ID. Product Type is one-hot encoded. The retained representation contains five operating variables and three type indicators. The two-stage stratified split targets 65%/15%/20% and yields 6,500 training, 1,499 validation and 2,001 test observations under the recorded implementation; the validation/test counts reflect rounding in the second split. Non-numeric or missing predictor values are replaced deterministically with zero, and StandardScaler parameters for all retained columns are fitted on the training partition only.

MetroPT3 observations are sorted by timestamp and divided at fixed row percentages: the first 55% form training, the next 20% validation, and the final 25% the test set, with integer truncation for the first two sizes. The boundaries are not selected by an event-aware algorithm. Training contains 834,321 observations from 1 February 2020, 00:00:00 to 29 May 2020, 05:49:30; validation contains 303,389 from 29 May 2020, 05:49:40 to 10 July 2020, 03:07:15; and testing contains 379,238 from 10 July 2020, 03:07:25 to 1 September 2020, 03:59:50. Their positive counts are 8,657, 19,675 and 1,622, respectively. Timestamp and label columns are excluded from predictors, every retained row is used, and scaling is fitted exclusively on training observations.

The split audit reports contiguous positive runs and recovered-interval membership separately from official physical-event identities. Chronological order prevents future observations from entering training or scaler fitting, but this single temporal holdout does not measure generalization across independent future episodes. Validation

**Table 3** Summary of the AI4I 2020 dataset and preprocessing used in this study.

| Attribute | Description |
|---|---|
| Dataset size | 10,000 observations |
| Machine descriptors | Product type and operating variables |
| Continuous variables | Air temperature, process temperature, rotational speed, torque, and tool wear |
| Failure indicators | TWF, HDF, PWF, OSF, and RNF |
| Target construction | Binary target obtained from the failure indicators; original failure-related fields removed before modelling |
| Categorical encoding | *Type* one-hot encoded |
| Scaling | StandardScaler fitted on all retained training columns only |
| Data split | Stratified random nominal 65%/15%/20%; actual counts 6,500/1,499/2,001 |
| Source | AI4I 2020 predictive maintenance dataset [22] |

**Table 4** Summary of the MetroPT3 dataset and preprocessing used in this study.

| Attribute | Description |
|---|---|
| Dataset contents | 1,516,948 observations from the 2020 compressor recording; 15 predictors |
| Representative variables | TP2, TP3, H1, DV_pressure, Reservoirs, Oil_temperature, Motor_current, COMP, DV_eletric, Towers, MPG, LPS, Pressure_switch, Oil_level, and Caudal_impulses |
| Failure label | Accepted processed `machine_failure`; recovered timestamp rule is compatible with this snapshot, with producing provenance unresolved |
| Temporal handling | Records ordered chronologically before partitioning |
| Timestamp usage | Used for ordering and removed before model input |
| Scaling | Fitted on the training period only |
| Data split | Fixed-row chronological 55%/20%/25%, without event-selected boundaries |
| Source | MetroPT3 compressor dataset [18] |

supports checkpoint, learning-rate, and threshold selection; test metrics do not enter these selection rules.

## 3.4 Multi-seed evaluation and client allocation

Seed identifiers label reproducible runs, rather than their count. The core and paired attribution comparisons use ten matched seeds (5–14); component, feature-order, loss and gradient studies use five (5–9); and federation, participation and learning-rate sensitivity studies use three (5–7). Each study compares its conditions within the same seed set. Model initialization and client allocation are repeated by seed, and AI4I stratification also changes; MetroPT3 retains the same chronological partitions. Identifiers 0–4 belong to the separate historical campaign and are not pooled with these estimates.

Federated clients are constructed from the training partition only. Three clients are used with target sample fractions

$$f = (0.40, 0.30, 0.30).$$

**Algorithm 2** Dirichlet label-non-IID client allocation

**Require:** Training labels $y \in \{0,1\}^n$; number of clients $K$=3; target fractions $f = (0.40, 0.30, 0.30)$; concentration $\alpha$=0.5; random seed $s$

**Ensure:** Disjoint client index sets $\{I_k\}_{k=1}^K$ with $|I_k| \approx f_k n$

Draw independently for each class $c \in \{0,1\}$:

$$p_c \sim \text{Dirichlet}(\alpha, \ldots, \alpha)$$

Determine target client sizes $n_k \leftarrow \text{round}(f_k n)$ for $k < K$; set $n_K \leftarrow n - \sum_{k<K} n_k$

**for** each class $c$ **do**

  Let $J_c = \{i : y_i = c\}$ and shuffle $J_c$ using seed $s$

  Split $J_c$ using cumulative boundaries $\lfloor |J_c| \sum_{j \le k} p_{c,j} \rfloor$, with last boundary $|J_c|$

  Append block $k$ to provisional client index set $I_k$

**end for**

Shuffle each provisional client index set $I_k$

Trim oversized clients to $n_k$ and concatenate excess indices into a pool

Shuffle the pool and fill client deficits sequentially in client order

**return** $\{I_k\}_{k=1}^K$

Class-specific proportions are sampled from a Dirichlet distribution with concentration parameter $\alpha = 0.5$. Provisional class allocations are rebalanced to the target client sizes, assigning each training observation exactly once. The same seeded allocation procedure is used across model families to provide matched client membership.

The generated partitions are audited using client-level positive-class rates and total-variation distance from the corresponding global training distribution. These diagnostics confirm genuine but modest label heterogeneity, particularly for the highly imbalanced MetroPT3 data. The experiments therefore evaluate controlled label heterogeneity and do not represent a maximally adversarial non-IID setting.

Algorithm 2 summarises the allocation procedure.

## 3.5 Centralised training

For each dataset and seed, CNN, BiLSTM, MLP, CNN–LSTM, and FusionNet Lite are trained independently on the complete centralized training partition. Adam optimization with learning rate $10^{-3}$ and binary cross-entropy loss is used in the core comparison, with the training batch size of 2048.

Centralized models are trained for up to 80 epochs with early stopping based on validation area under the receiver operating characteristic curve (AUROC). The best validation-performing weights are retained before final evaluation on the held-out test partition. Early stopping uses a patience of 12 epochs. Exact AUROC is computed from the complete validation predictions after each epoch, and only strict improvement replaces the saved best checkpoint, retaining the earliest maximum on ties. Final weights are saved separately at termination. Both checkpoints are evaluated and identified separately in the outputs; primary centralized–FedAvg comparisons use best-validation checkpoints in both modes. This aligns selection criteria while retaining

**Table 5** Core FedAvg configuration

| Component | Configuration |
|---|---|
| Aggregation algorithm | Synchronous FedAvg |
| Models | CNN, BiLSTM, MLP, CNN–LSTM, FusionNet Lite |
| Number of clients | 3 |
| Target client fractions | 40%, 30%, 30% |
| Client allocation | Dirichlet label-non-IID |
| Dirichlet concentration | $\alpha = 0.5$ |
| Communication rounds | 20 |
| Local epochs | 3 per round |
| Batch size | 2048 |
| Optimiser | Adam ($\eta = 10^{-3}$) |
| Loss | Binary cross-entropy |
| Aggregation | Sample-size weighted |
| Round-wise monitoring | Validation AUROC |
| Final evaluation | Held-out test partition |
| Checkpoint selection | Highest exact validation AUROC; earliest checkpoint on ties |
| Evaluation checkpoints | Best-validation and final round, reported separately |
| Optimizer state | Fresh Adam optimizer for every local client fit |

the distinct optimization budgets of centralized early stopping and 20 federated rounds with three local epochs.

## 3.6 FedAvg training

The core federated experiments simulate synchronous FedAvg [23] for all five model families on one compute host. For a given dataset and seed, the same client partition is reused across CNN, BiLSTM, MLP, CNN–LSTM, and FusionNet Lite so that architectural comparisons are made under an identical data allocation.

At the beginning of each communication round, the current global model is distributed to all three clients. Each client creates a fresh local model and Adam optimizer, initializes the model from the global weights, and performs three local epochs with learning rate $10^{-3}$, binary cross-entropy loss and batch size 2048. Clients return updated weights to the server, which combines them according to local sample count.

For client $k$ containing $n_k$ training observations, the aggregated parameters at round $t$ are

$$w^{(t)} = \sum_{k=1}^{K} \frac{n_k}{\sum_{j=1}^{K} n_j} w_k^{(t)}, \tag{1}$$

where $w_k^{(t)}$ denotes the locally updated parameters of client $k$. Twenty communication rounds are performed. Exact AUROC from the complete validation predictions is recorded after each aggregation, and strict improvement saves a best-validation checkpoint. Final-round weights are saved separately. Test predictions are excluded from round-wise monitoring and checkpoint selection. Best and final checkpoints are compared within the same fit.

Algorithm 3 provides the model-independent federated procedure.

**Algorithm 3** Sample-size-weighted FedAvg applied to each model family

**Require:** Model builder $\mathcal{B}$; rounds $R$=20; local epochs $E$=3; batch size $B$=2048; learning rate $\eta$=$10^{-3}$; client sets $\{I_k\}$ from Algorithm 2; training arrays $(X, y)$; validation data $(X^{\text{val}}, y^{\text{val}})$

Initialise global weights $w^{(0)}$ from a fresh $\mathcal{B}(\,)$
$a_{\text{best}} \leftarrow -\infty$
**for** $t = 1 \rightarrow R$ **do**
  $\mathcal{W} \leftarrow [\,], \mathcal{N} \leftarrow [\,]$
  **for** $k = 1 \rightarrow K$ **do**
    $n_k \leftarrow |I_k|$
    $m_k \leftarrow \mathcal{B}(\,)$; create a fresh Adam optimizer
    Set $m_k$ weights to $w^{(t-1)}$
    Train $m_k$ on $(X[I_k], y[I_k])$ for $E$ epochs with batch size $B$
    Append the resulting weights to $\mathcal{W}$
    Append $n_k$ to $\mathcal{N}$
  **end for**
  $w^{(t)} \leftarrow \sum_{k=1}^{K} \frac{n_k}{\sum_j n_j} \mathcal{W}[k]$
  Compute exact validation AUROC $a_t$ from the complete validation predictions
  **if** $a_t > a_{\text{best}}$ **then**
    $a_{\text{best}} \leftarrow a_t$; save $w_{\text{best}} \leftarrow w^{(t)}$
  **end if**
**end for**
$w_{\text{final}} \leftarrow w^{(R)}$
**return** $w_{\text{best}}$, $w_{\text{final}}$ and the validation AUROC trajectory

We intentionally retained FedAvg as the only aggregation method in this study. This isolates the effects of architecture, client heterogeneity, and optimization mode without introducing additional algorithm-dependent variation. Alternative aggregation and personalized methods are considered in the limitations and future-work discussion rather than being treated as completed experiments.

### 3.7 Controlled ablations and sensitivity analyses

Five architectural ablations remove SE attention, the stem residual, the mixer residual, or the complete mixer, or replace the two separable stem convolutions with standard Conv1D layers. Each variant is retrained under both optimization modes using diagnostic seeds $\{5, \ldots, 9\}$ and the same data partitions, checkpoint rule and training limits as its matched full-model control. Parameter counts and measured latency are reported separately because these changes alter model complexity. Feature-order experiments use the same seeds and compare the canonical ordering with its reversal and one fixed permutation generated with seed 17. Each ordering is applied consistently to training, validation and test features before retraining.

Federation sensitivity uses FusionNet Lite and MLP with seeds $\{5, 6, 7\}$, crossing $K \in \{3, 10\}$ with $\alpha \in \{0.1, 0.5, 10\}$. These arms use equal target client sizes at each $K$; their $K = 3$ control therefore differs in quantity allocation from the core

$(0.40, 0.30, 0.30)$ configuration. A separate $K = 10$, $\alpha = 0.5$ study samples half of the clients per round, then applies independent dropout probabilities of zero or 0.2. Aggregation is normalized by the training-sample count of the active clients; when all sampled clients drop out, the global weights remain unchanged. Realized participation and validation trajectories are retained.

For the active set $A_t$, the generalized aggregation is

$$w^{(t)} = \begin{cases} \sum_{k \in A_t} \frac{n_k}{\sum_{j \in A_t} n_j} w_k^{(t)}, & A_t \neq \varnothing, \\ w^{(t-1)}, & A_t = \varnothing. \end{cases} \tag{2}$$

Imbalance-loss comparisons use FusionNet Lite and MLP under FedAvg over diagnostic seeds $\{5, \ldots, 9\}$. Weighted binary cross-entropy uses class weights $N/(2N_c)$ calculated from the complete training partition and shared unchanged across clients. The focal-loss arm uses positive-class $\alpha = 0.25$, negative-class weight 0.75 and $\gamma = 2$; it is evaluated as specified, with no claim that it universally increases minority-class weight. Both alternatives retain a matched unweighted-BCE control and are evaluated under the same threshold policies.

Finally, all five model families are assessed at learning rates $\{3 \times 10^{-4}, 10^{-3}, 3 \times 10^{-3}\}$ on the same three sensitivity seeds. One learning rate is selected for each dataset, model and optimization mode by the mean best-validation AUROC across those matched seeds. Exact ties prefer $10^{-3}$, then the smaller rate. Test metrics do not enter this ranking. This bounded search assesses sensitivity to the shared default; it does not claim globally optimal model configurations.

Training-only diagnostics use the core FedAvg FusionNet Lite, and MLP fits at diagnostic seeds $\{5, \ldots, 9\}$, computing unweighted-BCE parameter gradients and input-channel derivatives at each fit's fixed best-validation global checkpoint from a deterministic batch of recorded training observations per client, with dropout disabled and no parameter updates. Parameter-gradient cosines compare client pairs and each client with the sample-count-weighted batch-gradient estimate; input derivatives are rescaled to per-example derivatives and summarized by feature and class. Pearson predictor correlations are computed from standardized training features once per distinct dataset and training split; these diagnostics describe associations rather than establish causal mechanisms.

## 3.8 Integrated Gradients and attribution agreement

IG [24] is applied to all five model families to compare centralized and federated attributions on fixed held-out inputs. For an input $\mathbf{x}$, baseline $\mathbf{x}'$ and model output function $F(\cdot)$, the attribution assigned to feature $i$ is

$$\mathrm{IG}_i(\mathbf{x}) = (x_i - x_i') \int_0^1 \frac{\partial F(\mathbf{x}' + \alpha(\mathbf{x} - \mathbf{x}'))}{\partial x_i} \, d\alpha. \tag{3}$$

The zero tensor in standardized feature space is used as the baseline. Input gradients are averaged with equal weights at 49 uniformly spaced interpolation points

**Algorithm 4** Class-specific Integrated Gradients

**Require:** Trained model $F$ from one family, used in inference mode; zero baseline $x'$ of shape $(1, F)$ for MLP or $(1, F, 1)$ otherwise; paired held-out batch $x$ of shape $(N, F)$ or $(N, F, 1)$ from one class, with $N \leq 128$; steps $m$=48

Convert $x$ and $x'$ to float32
Generate $m + 1$ uniformly spaced coefficients $\{\alpha_i\}_{i=0}^{m}$ on $[0, 1]$
$x'_{\text{batch}} \leftarrow \text{repeat}(x', N)$
Initialise $G \leftarrow 0$
**for** $i = 0 \rightarrow m$ **do**
    $x_{\alpha_i} \leftarrow x'_{\text{batch}} + \alpha_i(x - x'_{\text{batch}})$
    $g_i \leftarrow \nabla_{x_{\alpha_i}} \sum F(x_{\alpha_i})$
    $G \leftarrow G + g_i$
**end for**
$\bar{G} \leftarrow G/(m + 1)$
$\text{IG}(x) \leftarrow (x - x'_{\text{batch}}) \odot \bar{G}$
Average attributions over observations and remove any singleton channel dimension to obtain one value per feature
**return** attribution vector

corresponding to 48 steps. For each dataset and fresh core seed (5–14), up to 128 held-out observations are selected deterministically without replacement from each class. Identical observation IDs are used for centralized and FedAvg models and across all five model families. Attributions use best-validation checkpoints in inference mode, without updating the models. The class-specific sampling includes rare positives directly and is not weighted by population class prevalence.

For each model, seed, and test class, we compare signed-mean feature attributions using Spearman and Kendall correlations, sign agreement, top-three Jaccard overlap based on absolute signed-mean attribution, and sign-change counts. Mean absolute attributions and absolute deltas accompany the signed quantities. Magnitude-filtered sign agreement retains only features whose absolute signed-mean attribution is at least $10^{-6}$ in both optimization modes. Undefined correlations and empty magnitude-filtered comparisons remain unavailable rather than being replaced with zero. Statistics are summarized across the defined seed-level values, retaining their valid counts.

The archived FusionNet Lite analysis on the first 256 test observations remains a separate sampling protocol. The paired class-specific analysis covers all five model families. Attribution disagreement describes the evaluated models and inputs; it does not establish causal sensor effects or a mechanism of gradient starvation.

## 3.9 Evaluation metrics and statistical analysis

Predictive evaluation uses several complementary metrics because both datasets are imbalanced and accuracy alone cannot adequately characterize failure detection. The area under the receiver operating characteristic curve and area under the precision–recall curve (AUPRC) provide threshold-independent measures of ranking

**Algorithm 5** Validation-only threshold selection and held-out evaluation

**Require:** Fixed checkpoint predictions on validation and test partitions; declared threshold grids

1: **for** original-grid $F_1$, expanded-grid $F_1$ and expanded-grid $F_2$ policies **do**
2: Evaluate the policy objective on validation predictions only
3: Select the smallest threshold attaining its maximum
4: Apply that locked threshold to test predictions; retain precision, recall, $F_1$, $F_2$ and confusion counts
5: **end for**
6: Evaluate fixed 0.50 separately as a diagnostic, without selection
7: Retain test AUROC and average precision independently of threshold policy

performance, while $F_1$, $F_2$, and accuracy characterize threshold-dependent classification behavior. AUPRC and $F_2$ receive particular attention when interpreting rare-failure detection because they are more sensitive to positive-class performance than overall accuracy.

AUPRC is computed using the average-precision implementation in scikit-learn. Four threshold policies are evaluated separately for each checkpoint. The original policy maximizes validation $F_1$ over 200 evenly spaced values from 0.05 to 0.95. An expanded grid adds 0.001, 0.005 and increments of 0.01 from 0.01 to 0.30 to the original grid. Maximizing validation $F_1$ or $F_2$ on this same expanded grid separates the effect of the grid from that of the objective. The fourth policy fixes the threshold at 0.50 as a diagnostic control. Ascending-grid ties retain the first maximum, and test labels never select a threshold. Test probabilities greater than or equal to the locked threshold are classified as positive.

Expanded validation-$F_2$ selection supplies the primary $F_2$ estimates; fixed 0.50 is retained only for the diagnostic comparison.

For $\beta \in \{1, 2\}$, the count-based objective is

$$F_\beta = \frac{(1+\beta^2)\mathrm{TP}}{(1+\beta^2)\mathrm{TP} + \beta^2\mathrm{FN} + \mathrm{FP}}, \tag{4}$$

with zero returned when the denominator is zero. Validation selection is

$$\tau_\beta^\star = \operatorname{minarg\,max}_{\tau \in \mathcal{G}} F_\beta\big(y^{\mathrm{val}}, \mathbf{1}[p^{\mathrm{val}} \geq \tau]\big), \tag{5}$$

where $\mathcal{G}$ is the declared grid for that policy.

We evaluated computational efficiency using trainable parameter count and measured amortized batch inference latency. The Energy Efficiency Index (EEI) used in this study is defined as

$$\mathrm{EEI} = P \times L_{\mathrm{batch}}, \tag{6}$$

where $P$ is the number of trainable parameters and $L_{\mathrm{batch}}$ is measured latency in milliseconds per sample; EEI therefore has units of parameter·ms/sample. Lower EEI

values indicate a smaller combined parameter–latency cost. EEI is therefore treated as an experimental efficiency proxy instead of a direct measurement of electrical energy consumption.

For the fixed-batch EEI calculation, elapsed batch milliseconds are divided by 256 observations. Timing uses one warm-up followed by five calls per checkpoint and seed to the host `predict` API at input sizes of one and 256 observations, for best and final checkpoints of the core models and structural ablations. Each elapsed call is divided by its actual input size. Repetitions are summarized within a seed before means and sample standard deviations are formed across ten core seeds or five ablation seeds; they are not independent seed repetitions. These timings include API and execution overhead on the recorded GPU host. GPU-board telemetry is retained with its sampled coverage and is not attributed to individual models as electrical energy.

Fresh core results are summarized as mean and sample standard deviation over ten matched seeds. Sixteen declared comparisons contrast FusionNet Lite with MLP and BiLSTM for AUPRC and $F_2$ across two datasets and two optimization modes. They use best-validation checkpoints, AUPRC from the original-policy row and $F_2$ from expanded validation-$F_2$ selection. Independent recomputation reproduces the saved two-sided Wilcoxon probabilities using exact calculation for contrasts without zero paired differences and the normal approximation with Wilcox exclusion for the two contrasts containing zero differences. Holm's procedure adjusts probabilities across the complete family of 16 contrasts. Paired mean and median differences and descriptive 95% $t$ intervals accompany the tests. These seeds characterize variability under the specified protocol; they are not independent industrial cohorts or future MetroPT3 periods.

## 3.10 Reproducibility and consistency auditing

The experimental workflow includes a reproducibility package that preserves the connection between preprocessing, training, seed-level outputs, summary tables, figures, and explainability results. The fresh campaign comprises 596 fits and 4,768 evaluation rows, with two checkpoints and four threshold policies per fit. Evaluation rows are not independent training repetitions. The historical five-seed reference remains separate.

Consistency checks test coverage, source identity, partition separation, checkpoint selection, metric/confusion identities, attribution pairing and agreement between numerical exports. Completed records can be reused only after their saved configuration and artifact identities are verified. Such checks establish consistency of the saved computational record; they do not independently reproduce historical training or resolve missing annotation provenance.

The complete implementation, environment specification, deterministic runner, reference result artifacts, and consistency-audit procedures are available through the accompanying Code Ocean reproducibility capsule. The reproducible workflow covers preprocessing, centralized and federated training, seed-level evaluation, attribution analysis, and generation of the associated result artifacts

within a defined computational environment. See the Data and Code Availability statement for the capsule identifier, access details, and dataset availability information. Four standalone supplementary CSVs provide predictive evaluation records (`full_canonical_metrics.csv`), seed-level attribution agreement (`full_ig_agreement_seeds.csv`), planned paired statistics (`significance.csv`) and realized client distributions (`full_client_heterogeneity.csv`). Full execution additionally requires the accepted input datasets, the recorded software environment and suitable compute resources; the experimental implementation generates the remaining diagnostic reports.

***Computational verification.***

Before the larger GPU campaign, OpenAI Codex (GPT-6.1 Sol Ultra) assisted with code interpretation, debugging and checks of partition separation, validation-only selection and saved-artifact consistency. Small local engineering checks preceded allocation-intensive centralized and federated runs, with the aim of limiting avoidable failed computations while preserving the established codebase. Codex also assisted with extracting the distributable implementation. These checks support verification without guaranteeing the absence of every error or leakage route.

The campaign used one NVIDIA A100-SXM4-80GB GPU and 32 Slurm CPUs per allocation on Ngarrgu Tindebeek within the OzSTAR national facility. Allocated memory increased from 128 to 256 GiB for the final continuation. The 32 CPUs provided headroom for data handling and TensorFlow operations; model fits and simulated-client updates ran sequentially. Successful training-routine intervals totalled 30.28 h, while the initial-allocation-to-first-full-export window was approximately 43 h, including pauses and restarts (Note: Neither value measures continuous GPU use).

# 4 Results

The main comparison comprises ten matched seeds (5–14) for five model families under centralized and FedAvg optimization on AI4I 2020 and MetroPT3. Both modes use checkpoints selected by exact validation AUROC; threshold-independent ranking metrics and threshold-dependent failure-detection metrics are reported separately. Additional component, feature-order, federation, loss and learning-rate analyses retain their own matched seed subsets. The earlier five-seed outputs are not pooled with these estimates.

Unless stated otherwise, results use best-validation checkpoints, and $F_2$ uses the expanded validation-$F_2$ policy defined in Sect. 3.9. Values are means and sample standard deviations across the stated seed set. The analysis examines the concerns identified in Sect. 2: client heterogeneity, matched model comparisons, minority-class detection, computational cost and quantitative explanation agreement.

## 4.1 Observed client heterogeneity

The three-client core allocation produces label-distribution differences as well as unequal client sizes. Across ten seeds, the seed-wise mean client-to-global total-variation distance ranges from 0.0130 to 0.0424 on AI4I and from 0.0043 to 0.0126 on

MetroPT3. Individual client positive rates range from 0 to 0.1041 and from 0 to 0.0310, respectively; two AI4I client partitions and one MetroPT3 client partition contain no positive observations.

These values describe controlled label heterogeneity but do not capture every form of industrial non-IID data. In particular, low total-variation distance can coexist with a client containing no failures because the global failure prevalence is small. The complete client counts and realized positive rates are therefore reported alongside the target allocation parameters in `full_client_heterogeneity.csv`. This interpretation follows the allocation audit in Sect. 3.4; a small distance under rare failures does not establish that every client observes both classes.

## 4.2 Predictive performance across training regimes

Table 6 reports the ten-seed comparison under matched validation checkpoint selection. On AI4I, centralized FusionNet Lite attains AUROC $0.7939 \pm 0.1701$, AUPRC $0.2319 \pm 0.1528$, and test $F_2$ $0.3617 \pm 0.1279$ using expanded validation-$F_2$ thresholds. MLP has a larger mean AUPRC, $0.2720 \pm 0.2770$, whereas the large dispersions prevent a narrow ranking of the two models from the means alone. FusionNet Lite uses 1,420 trainable parameters compared with 3,201 for MLP, but parameter count does not remove this uncertainty in predictive performance.

Under FedAvg, AI4I FusionNet Lite achieves AUROC $0.8289 \pm 0.0621$, AUPRC $0.2250 \pm 0.1189$, and $F_2$ $0.3508 \pm 0.0908$. Its mean AUPRC and $F_2$ exceed those of the evaluated federated baselines under this fixed learning-rate protocol, although neither Lite–MLP primary-metric contrast remains significant after multiplicity adjustment. The comparison therefore separates a descriptive advantage in this configuration from evidence of general architectural superiority.

On the later MetroPT3 test interval, centralized FusionNet Lite obtains AUROC $0.9539 \pm 0.0250$, AUPRC $0.3110 \pm 0.1380$, and $F_2$ $0.3537 \pm 0.1026$. Under FedAvg, AUROC remains high at $0.9351 \pm 0.0462$, while AUPRC decreases to $0.2197 \pm 0.1971$ and $F_2$ to $0.1566 \pm 0.2044$. BiLSTM and CNN–LSTM attain larger federated mean AUPRC values, $0.2445 \pm 0.2289$ and $0.2453 \pm 0.1317$, and larger $F_2$ values, $0.2095 \pm 0.2680$ and $0.2021 \pm 0.1675$. Thus, the strongest mean AUROC does not identify the strongest rare-failure detector.

The CNN and MLP remain weak MetroPT3 detectors under the core configuration, and the threshold analyses below show that lowering the decision threshold does not rescue every architecture. Figure 3 summarizes ranking discrimination, while Figs. 4 and 5 show AUPRC and $F_2$. Reading them together separates ranking performance from thresholded failure detection.

**Table 6** Core predictive performance across ten matched seeds.

| Dataset | Model | Mode | Seeds | AUROC | AUPRC | $F_2$ (val $F_1$) | $F_2$ (val $F_2$) |
|---|---|---|---|---|---|---|---|
| ai4i | BiLSTM | central | 10 | $0.70855 \pm 0.217179$ | $0.200492 \pm 0.116851$ | $0.261707 \pm 0.130191$ | $0.304636 \pm 0.106051$ |
| ai4i | BiLSTM | fedavg | 10 | $0.683808 \pm 0.0526986$ | $0.107272 \pm 0.0479948$ | $0.126433 \pm 0.0695281$ | $0.214796 \pm 0.0261693$ |
| ai4i | CNN | central | 10 | $0.547944 \pm 0.231336$ | $0.130443 \pm 0.182075$ | $0.197116 \pm 0.168623$ | $0.237325 \pm 0.144528$ |
| ai4i | CNN | fedavg | 10 | $0.620467 \pm 0.169229$ | $0.121938 \pm 0.0772481$ | $0.195061 \pm 0.0678817$ | $0.233096 \pm 0.0724771$ |
| ai4i | CNN_LSTM | central | 10 | $0.638009 \pm 0.248895$ | $0.163 \pm 0.143539$ | $0.259761 \pm 0.108139$ | $0.278689 \pm 0.124432$ |
| ai4i | CNN_LSTM | fedavg | 10 | $0.421587 \pm 0.167595$ | $0.0493785 \pm 0.0495181$ | $0.0597213 \pm 0.102823$ | $0.182953 \pm 0.0547211$ |
| ai4i | FusionNetLite | central | 10 | $0.79389 \pm 0.170073$ | $0.231931 \pm 0.152799$ | $0.322911 \pm 0.129397$ | $0.361708 \pm 0.1279$ |
| ai4i | FusionNetLite | fedavg | 10 | $0.828869 \pm 0.062124$ | $0.225024 \pm 0.118942$ | $0.255478 \pm 0.158385$ | $0.350832 \pm 0.0908306$ |
| ai4i | MLP | central | 10 | $0.702733 \pm 0.223238$ | $0.271987 \pm 0.277016$ | $0.337674 \pm 0.237761$ | $0.35175 \pm 0.207476$ |
| ai4i | MLP | fedavg | 10 | $0.736823 \pm 0.0756323$ | $0.208608 \pm 0.07937$ | $0.263495 \pm 0.0886527$ | $0.313113 \pm 0.0750099$ |
| metro | BiLSTM | central | 10 | $0.941623 \pm 0.0800104$ | $0.262004 \pm 0.22753$ | $0.0987429 \pm 0.121787$ | $0.226734 \pm 0.247073$ |
| metro | BiLSTM | fedavg | 10 | $0.928871 \pm 0.0924644$ | $0.244455 \pm 0.228858$ | $0.0287854 \pm 0.0610559$ | $0.209471 \pm 0.26801$ |
| metro | CNN | central | 10 | $0.439735 \pm 0.106551$ | $0.0113989 \pm 0.0161675$ | $0 \pm 0$ | $0.000152952 \pm 0.000483677$ |
| metro | CNN | fedavg | 10 | $0.338122 \pm 0.0809429$ | $0.00703506 \pm 0.0125686$ | $0 \pm 0$ | $0 \pm 0$ |
| metro | CNN_LSTM | central | 10 | $0.815762 \pm 0.195521$ | $0.211478 \pm 0.19719$ | $0.171005 \pm 0.205942$ | $0.283666 \pm 0.217292$ |
| metro | CNN_LSTM | fedavg | 10 | $0.885016 \pm 0.126702$ | $0.245297 \pm 0.131714$ | $0.0573896 \pm 0.12503$ | $0.202112 \pm 0.167543$ |
| metro | FusionNetLite | central | 10 | $0.953856 \pm 0.0250161$ | $0.310968 \pm 0.138045$ | $0.0966051 \pm 0.106609$ | $0.353707 \pm 0.102586$ |
| metro | FusionNetLite | fedavg | 10 | $0.935055 \pm 0.0461643$ | $0.219683 \pm 0.197102$ | $0.0422504 \pm 0.0995371$ | $0.156635 \pm 0.204369$ |
| metro | MLP | central | 10 | $0.42698 \pm 0.200678$ | $0.00621927 \pm 0.00612894$ | $0 \pm 0$ | $7.25163 \times 10^{-5} \pm 0.000229317$ |
| metro | MLP | fedavg | 10 | $0.506334 \pm 0.15951$ | $0.00697745 \pm 0.00556328$ | $0 \pm 0$ | $0.000347717 \pm 0.000587683$ |

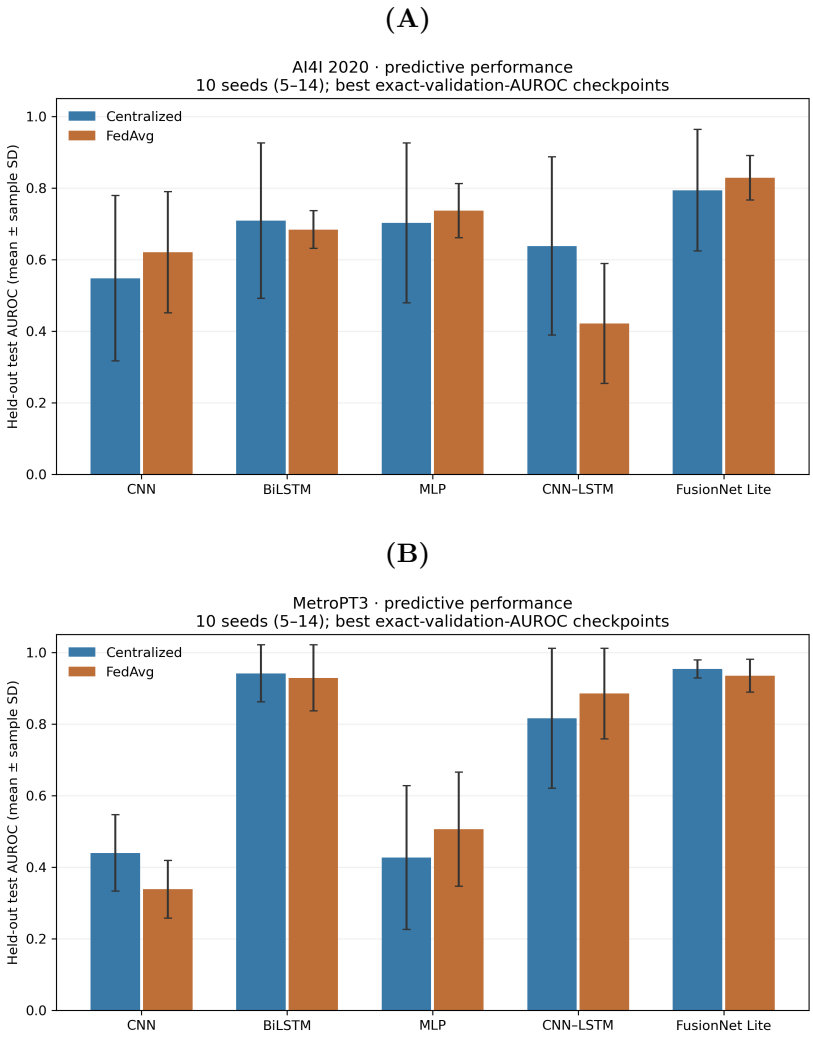
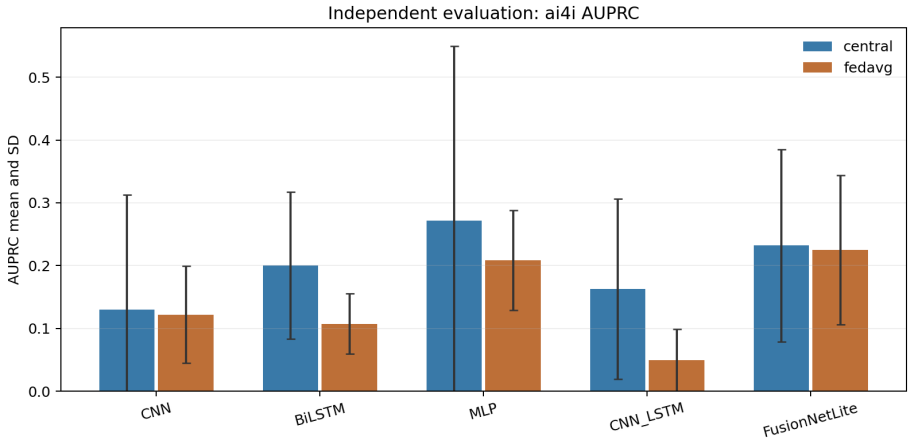



**Fig. 3** Test AUROC on AI4I (A) and MetroPT3 (B) across models and training modes.

**(A)**




**(B)**

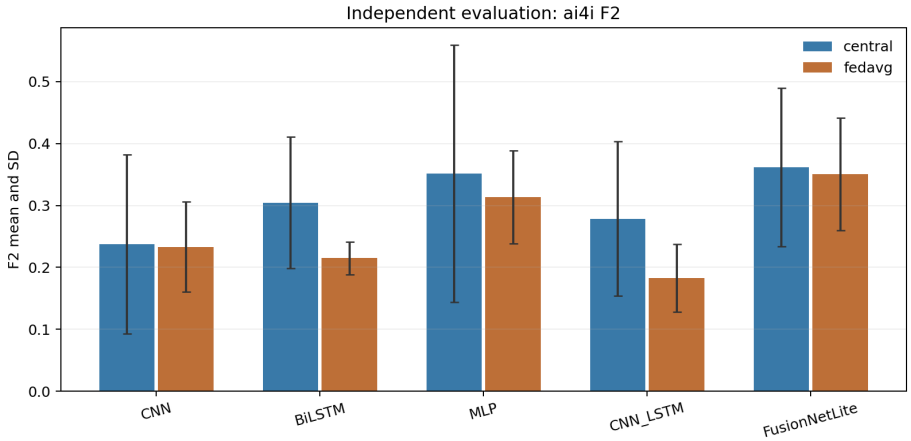



**Fig. 4** AI4I test AUPRC (A) and $F_2$ (B) across models and training modes.

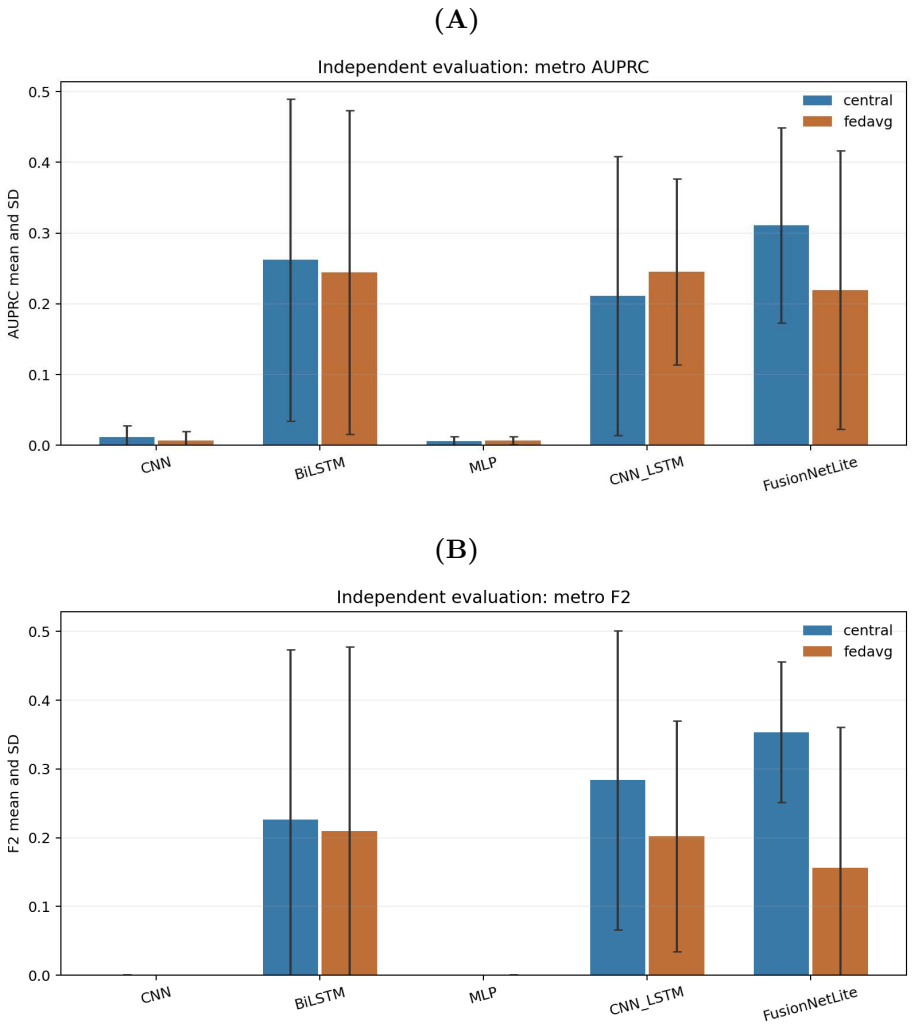



**Fig. 5** MetroPT3 test AUPRC (A) and $F_2$ (B) across models and training modes.

## 4.3 Statistical evidence and inferential limits

The statistical comparison uses ten matched seeds and a predeclared family of 16 Lite–MLP or Lite–BiLSTM contrasts across the two datasets, two training modes and two primary metrics. Four contrasts remain below $p = 0.05$ after Holm adjustment, each with adjusted $p = 0.03125$: AI4I FedAvg $F_2$ against BiLSTM, MetroPT3 centralized AUPRC and $F_2$ against MLP, and MetroPT3 FedAvg AUPRC against MLP. Their mean paired differences, defined as FusionNet Lite minus comparator, are 0.1360, 0.3047, 0.3536 and 0.2127, respectively.

The remaining comparisons do not support adjusted significance. In particular, FusionNet Lite does not differ significantly from MLP in either primary metric on AI4I, and neither MetroPT3 FedAvg primary-metric contrast against BiLSTM is significant. The larger campaign therefore supports several specific differences without establishing uniform superiority over the baseline families.

The complete raw and adjusted $p$-values and mean paired-difference intervals are supplied in `significance.csv`. Repeated seeds characterize variation within the specified datasets and partitioning protocols; AI4I stratified partitions also change with seed, whereas MetroPT3 retains one temporal test interval. These repetitions do not represent independent industrial sites or ten independent future test periods. The signed-rank tests and the descriptive $t$ intervals for mean differences also address different aspects of the paired results. The comparisons use the matched checkpoint and operating-point definitions in Sect. 3.9, rather than selecting a favorable test policy separately for each contrast.

## 4.4 FedAvg optimization dynamics

Figure 6 shows round-wise validation trajectories; Fig. 7 separately compares validation and test AUROC at the selected and final checkpoints. AI4I validation AUROC continues to improve over much of the 20-round budget; the trajectory alone therefore does not establish convergence. MetroPT3 validation AUROC remains high, while mean test AUROC is $0.9351 \pm 0.0462$ for the validation-selected checkpoint and $0.8975 \pm 0.0825$ for the final checkpoint. The distinction between validation monitoring and held-out checkpoint evaluation is essential to interpreting this gap.

Lines, shading and bars show means and sample SD across ten seeds. The round plots contain validation values only; the four-bar plots distinguish validation and test evaluations at best and final checkpoints. The checkpoint comparison follows Sects. 3.5 and 3.6: both states come from the same fit, and neither is selected by its test result. Paired checkpoint records are supplied in `full_canonical_metrics.csv`. These trajectories characterize 20 communication rounds with FedAvg; they do not establish the outcome of longer training or another aggregation rule.

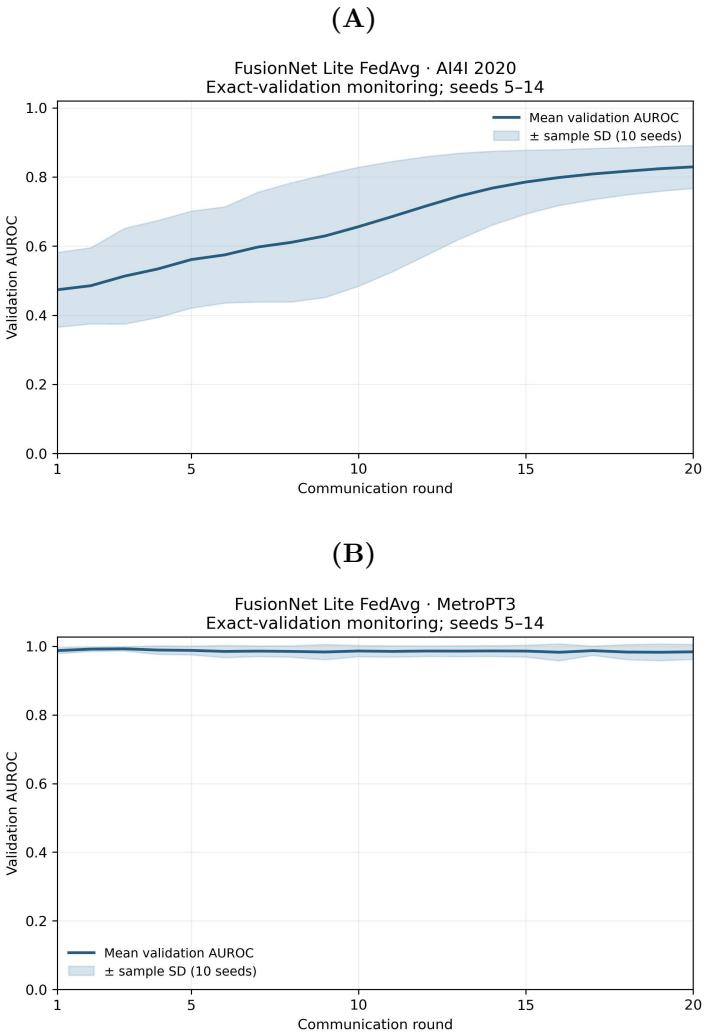
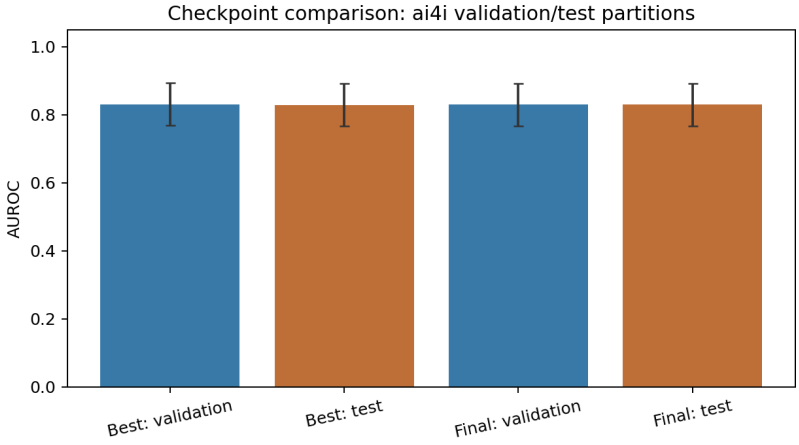



**Fig. 6** FusionNet Lite FedAvg validation AUROC over 20 rounds on AI4I (A) and MetroPT3 (B).

**(A)**




**(B)**

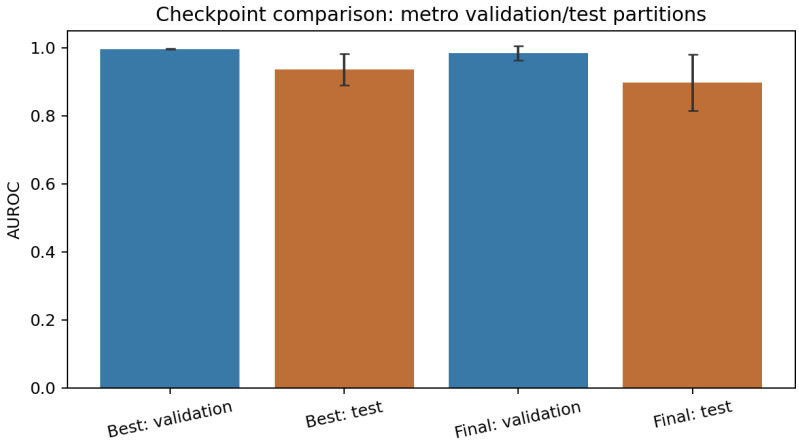



**Fig. 7** FusionNet Lite FedAvg validation and test AUROC at selected and final checkpoints on AI4I (A) and MetroPT3 (B).

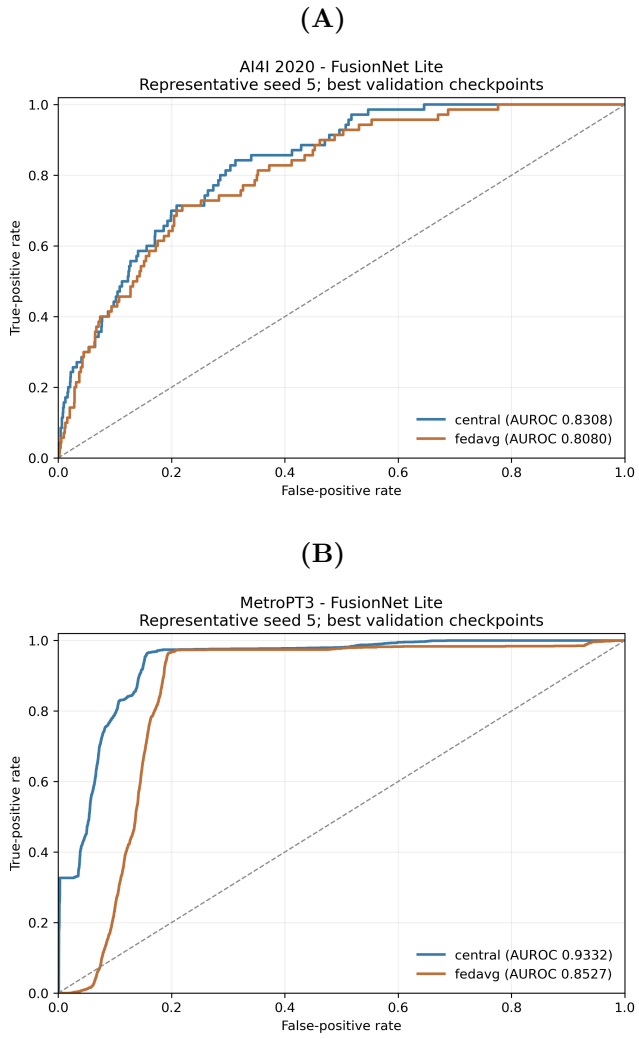



**Fig. 8** Seed-5 FusionNet Lite ROC curves on AI4I (A) and MetroPT3 (B) at validation-selected checkpoints.

## 4.5 Representative ROC behavior

Figure 8 illustrates the saved seed-5 predictions at validation-selected checkpoints. AI4I AUROC is 0.8308 centrally and 0.8080 under FedAvg; the MetroPT3 values are 0.9332 and 0.8527. These curves illustrate one seed and do not replace the ten-seed estimates in Table 6. In particular, the AI4I central–FedAvg ordering for seed 5 differs from the ordering of the aggregate means.

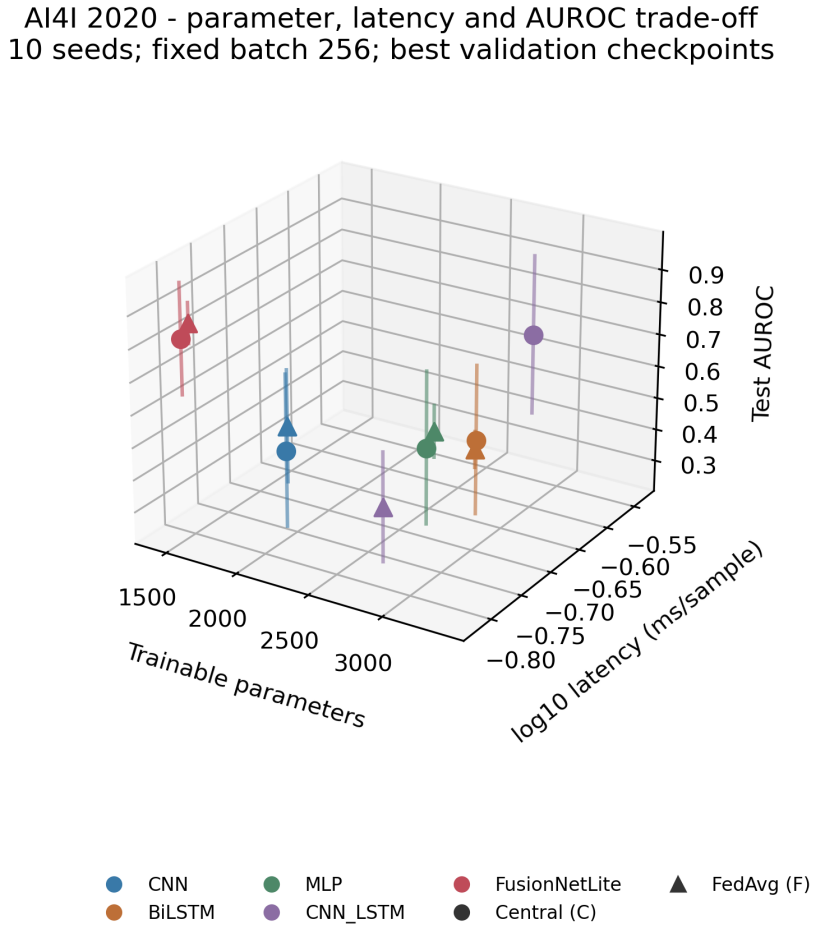



**Fig. 9** AI4I parameter count, $\log_{10}$ mean latency and mean test AUROC; vertical bars show AUROC SD.

## 4.6 Performance–complexity trade-off

FusionNet Lite retains 1,420 trainable parameters on both datasets, compared with 2,177–3,393 for the AI4I baselines and 2,753–3,713 for the MetroPT3 baselines. Figures 9 and 10 relate those counts to measured latency and held-out AUROC. On AI4I, FusionNet Lite combines the smallest parameter count with the largest mean AUROC in each training mode under the core protocol, although AUPRC and $F_2$ do not support uniform superiority.

FusionNet Lite also has the largest mean MetroPT3 AUROC in each mode, while the federated BiLSTM and CNN–LSTM have larger mean AUPRC and $F_2$. The three-dimensional views therefore describe a parameter–latency–AUROC relationship but do not capture every aspect of failure detection. Circles denote centralized models

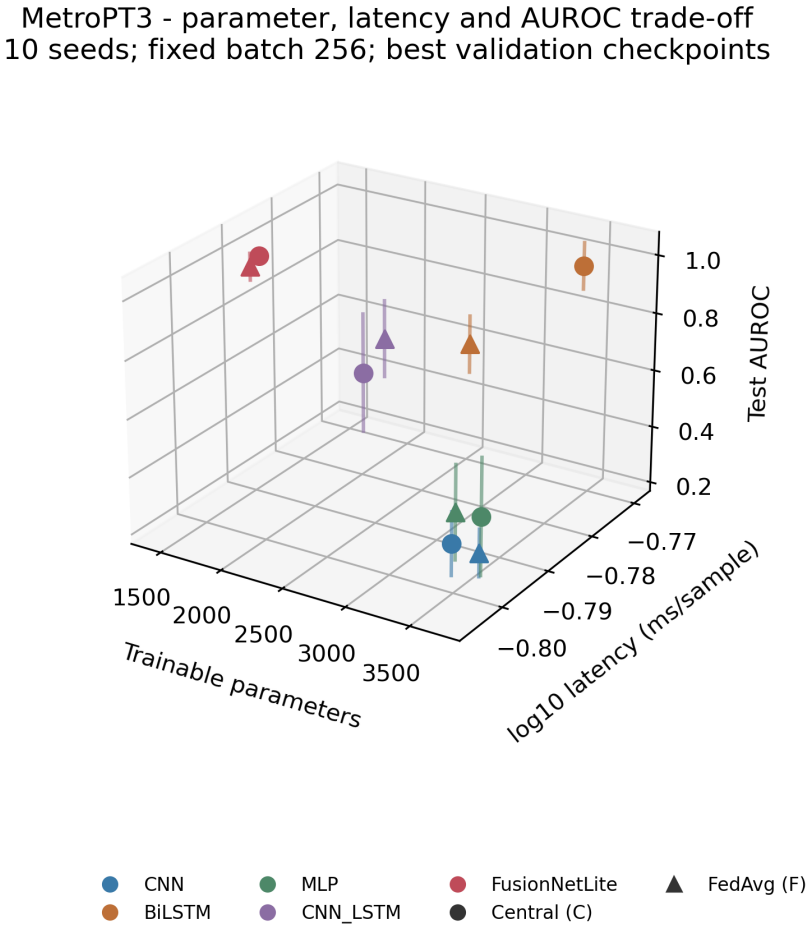



**Fig. 10** MetroPT3 parameter count, $\log_{10}$ mean latency and mean test AUROC; vertical bars show AUROC SD.

and triangles denote FedAvg models; only AUROC dispersion is plotted. Latency is the mean amortized time in ms/sample, shown on a $\log_{10}$ axis using the fixed-256 measurement protocol.

## 4.7 Efficiency proxy

Table 7 reports trainable parameters, measured latency, the parameter–latency index, serialized checkpoint size and training-routine time. With a fixed 256-sample GPU batch, FusionNet Lite has the smallest mean EEI in both modes on both datasets: $226.24 \pm 12.93$ and $232.24 \pm 17.15$ parameter·ms/sample on AI4I, and $232.08 \pm 13.07$ and $231.08 \pm 9.89$ parameter·ms/sample on MetroPT3, for centralized and FedAvg

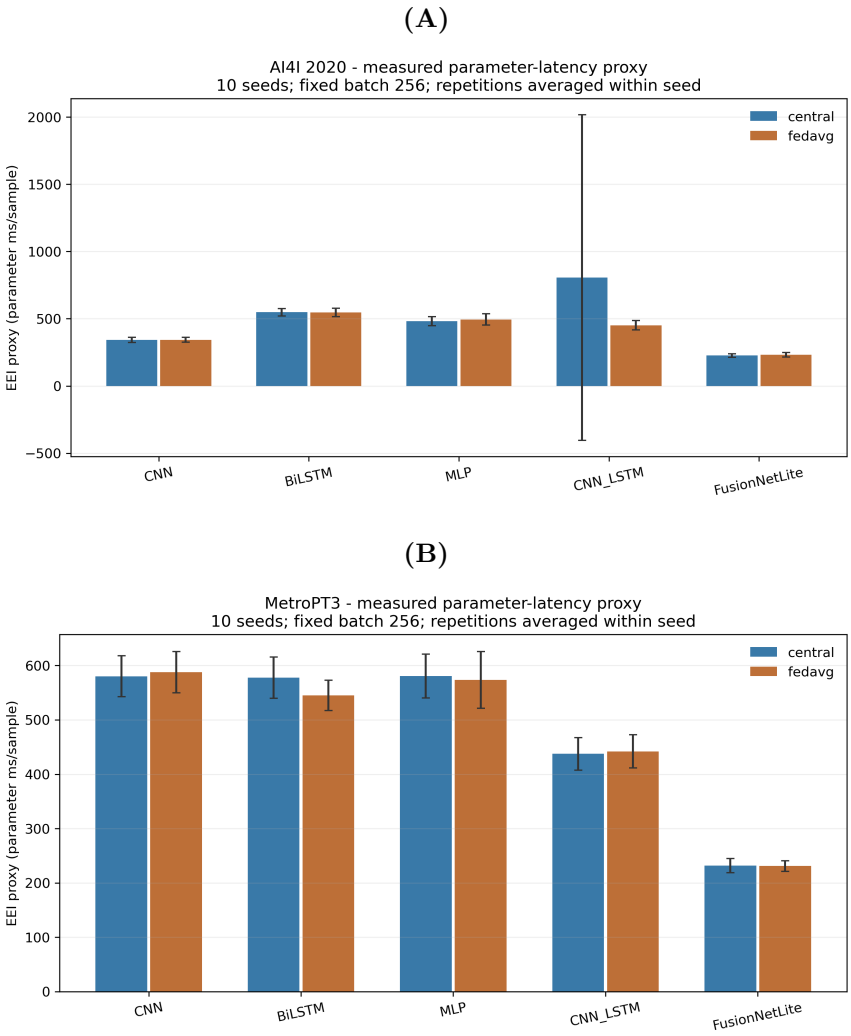



**Fig. 11** Parameter–latency proxy on AI4I (A) and MetroPT3 (B); bars show mean and sample SD.

models respectively. These values use the same batch-size measurement procedure for every model and dataset.

Figure 11 shows the mean and sample SD of this proxy. Figure 12 shows the same means by model and training mode on a $\log_{10}$ scale, without uncertainty bars; it complements the same measurements.

The parameter advantage does not imply minimum inference latency. On MetroPT3, centralized MLP averages $0.1591 \pm 0.0111$ ms/sample, compared with $0.1634 \pm 0.0092$ for FusionNet Lite; their EEI means are 580.66 and 232.08 parameter·ms/sample because MLP has 3,649 parameters. Likewise, the saved HDF5 file for FusionNet Lite is 56.375 KiB, compared with 33.523 KiB for the MetroPT3

**Table 7** Model resources and measured parameter–latency cost across ten seeds.

| Dataset | Model | Mode | Seeds | Parameters | Latency (ms/sample) | EEI (parameter ms/sample) | Weights (KiB) | Train/checkpoint time (s) |
|---|---|---|---|---|---|---|---|---|
| ai4i | BiLSTM | central | 10 | 3393 | $0.161534 \pm 0.00818671$ | $548.085 \pm 27.7775$ | $34.2578 \pm 0$ | $7.52146 \pm 2.60197$ |
| ai4i | BiLSTM | fedavg | 10 | 3393 | $0.160884 \pm 0.00919595$ | $545.879 \pm 31.2019$ | $34.2578 \pm 0$ | $259.678 \pm 7.01887$ |
| ai4i | CNN | central | 10 | 2177 | $0.157125 \pm 0.0090567$ | $342.062 \pm 19.7164$ | $26.6641 \pm 0$ | $2.14127 \pm 1.46477$ |
| ai4i | CNN | fedavg | 10 | 2177 | $0.157827 \pm 0.00791739$ | $343.589 \pm 17.2362$ | $26.6641 \pm 0$ | $65.2794 \pm 1.82898$ |
| ai4i | CNN_LSTM | central | 10 | 2753 | $0.292836 \pm 0.439386$ | $806.176 \pm 1209.63$ | $31.8516 \pm 0$ | $5.00107 \pm 2.60593$ |
| ai4i | CNN_LSTM | fedavg | 10 | 2753 | $0.164027 \pm 0.0125203$ | $451.566 \pm 34.4684$ | $31.8516 \pm 0$ | $163.377 \pm 5.23216$ |
| ai4i | FusionNetLite | central | 10 | 1420 | $0.159324 \pm 0.00910498$ | $226.24 \pm 12.9291$ | $56.375 \pm 0$ | $8.99221 \pm 2.62925$ |
| ai4i | FusionNetLite | fedavg | 10 | 1420 | $0.163548 \pm 0.0120766$ | $232.238 \pm 17.1488$ | $56.375 \pm 0$ | $248.849 \pm 6.3776$ |
| ai4i | MLP | central | 10 | 3201 | $0.150313 \pm 0.0105392$ | $481.15 \pm 33.736$ | $31.7734 \pm 0$ | $3.64657 \pm 2.36472$ |
| ai4i | MLP | fedavg | 10 | 3201 | $0.154481 \pm 0.0130429$ | $494.494 \pm 41.7502$ | $31.7734 \pm 0$ | $74.85 \pm 2.29553$ |
| metro | BiLSTM | central | 10 | 3393 | $0.170278 \pm 0.011147$ | $577.752 \pm 37.8217$ | $34.2578 \pm 0$ | $42.1793 \pm 13.081$ |
| metro | BiLSTM | fedavg | 10 | 3393 | $0.160696 \pm 0.00822841$ | $545.241 \pm 27.919$ | $34.2578 \pm 0$ | $355.258 \pm 9.17907$ |
| metro | CNN | central | 10 | 3713 | $0.156253 \pm 0.0101635$ | $580.167 \pm 37.7371$ | $32.6641 \pm 0$ | $14.6798 \pm 5.68748$ |
| metro | CNN | fedavg | 10 | 3713 | $0.158341 \pm 0.0102004$ | $587.922 \pm 37.8742$ | $32.6641 \pm 0$ | $112.427 \pm 2.39596$ |
| metro | CNN_LSTM | central | 10 | 2753 | $0.158914 \pm 0.010822$ | $437.49 \pm 29.793$ | $31.8516 \pm 0$ | $23.3597 \pm 5.42477$ |
| metro | CNN_LSTM | fedavg | 10 | 2753 | $0.160501 \pm 0.0110962$ | $441.86 \pm 30.5477$ | $31.8516 \pm 0$ | $235.991 \pm 4.81673$ |
| metro | FusionNetLite | central | 10 | 1420 | $0.163438 \pm 0.00920612$ | $232.082 \pm 13.0727$ | $56.375 \pm 0$ | $57.4036 \pm 10.7141$ |
| metro | FusionNetLite | fedavg | 10 | 1420 | $0.162735 \pm 0.00696722$ | $231.084 \pm 9.89346$ | $56.375 \pm 0$ | $437.125 \pm 8.34301$ |
| metro | MLP | central | 10 | 3649 | $0.159128 \pm 0.011124$ | $580.66 \pm 40.5913$ | $33.5234 \pm 0$ | $18.8657 \pm 6.10408$ |
| metro | MLP | fedavg | 10 | 3649 | $0.157158 \pm 0.0142538$ | $573.468 \pm 52.012$ | $33.5234 \pm 0$ | $121.098 \pm 1.85571$ |

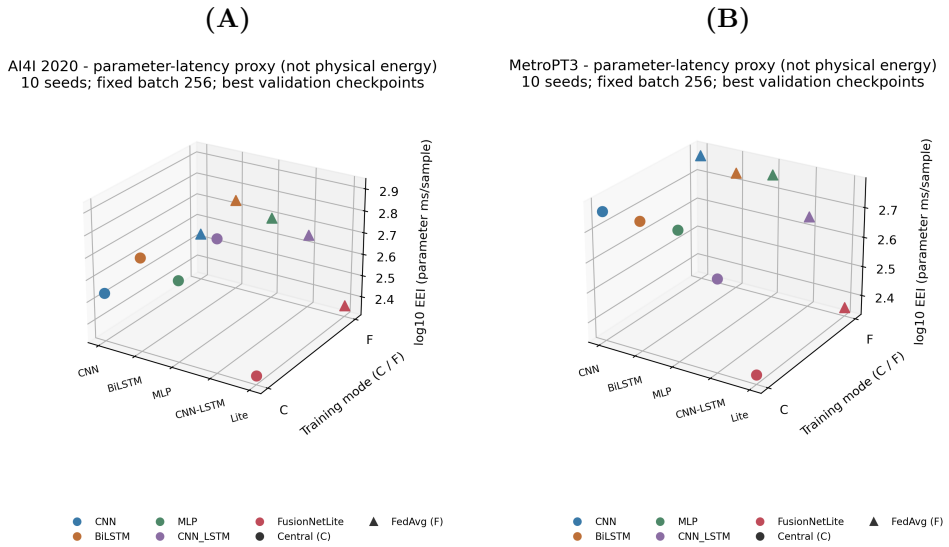



**Fig. 12** Model-by-mode views of $\log_{10}$ mean EEI on AI4I (A) and MetroPT3 (B).

MLP. Parameter count and serialized file size measure different costs, so the smaller parameter budget should not be presented as an unconditional storage advantage.

The fixed-batch timings describe throughput in the recorded GPU environment and include host/API overhead; they do not measure industrial-controller response times. Each seed reports the mean of five repetitions after warm-up. The reported standard deviation is calculated across seeds, not across repetitions treated as independent models. EEI remains a parameter–latency proxy; neither it nor the available stage-level GPU-board telemetry establishes per-model electrical energy, battery life or industrial edge-device performance.

## 4.8 Attribution agreement under federation

Table 8 compares central–FedAvg IG agreement for all five model families using class-stratified, paired held-out samples. For FusionNet Lite on AI4I, mean Spearman correlation is $0.5952 \pm 0.4185$ for non-failure observations and $0.7952 \pm 0.1151$ for failures; raw sign agreement is 0.8000 and 0.8875, respectively. The class-specific results therefore reveal different degrees of explanatory agreement that would be obscured by averaging an unstratified test prefix. Figures 13 and 14 show Spearman correlation and raw sign agreement, with mean and sample SD over defined seed-level values. Each class uses 128 paired observations per seed, except AI4I failures, for which 70 are available.

MetroPT3 FusionNet Lite has weaker and more variable agreement: Spearman correlation is $0.2407 \pm 0.2520$ for class 0 and $0.2754 \pm 0.5012$ for class 1, with raw sign agreement of 0.5533 and 0.5800. The mean rank correlations are positive, but their dispersion, top-3 overlaps of 0.28 and 0.37, and frequent sign changes show that the explanatory representation is not preserved uniformly.

**Table 8** Class-specific central–FedAvg IG agreement across ten matched seeds.

| Dataset | Model | Class | Paired samples/seed | Spearman | Raw sign agreement | Magnitude-filtered sign | Top-3 overlap |
|---|---|---|---|---|---|---|---|
| ai4i | BiLSTM | 0 | 128–128 | $0.347619 \pm 0.450805$ ($n = 10$) | $0.5625 \pm 0.237537$ ($n = 10$) | $0.5625 \pm 0.237537$ ($n = 10$) | $0.5 \pm 0.365148$ ($n = 10$) |
| ai4i | BiLSTM | 1 | 70–70 | $0.461905 \pm 0.451084$ ($n = 10$) | $0.725 \pm 0.210819$ ($n = 10$) | $0.725 \pm 0.210819$ ($n = 10$) | $0.77 \pm 0.3093$ ($n = 10$) |
| ai4i | CNN | 0 | 128–128 | $0.535714 \pm 0.537607$ ($n = 10$) | $0.7875 \pm 0.212867$ ($n = 10$) | $0.7875 \pm 0.212867$ ($n = 10$) | $0.59 \pm 0.307137$ ($n = 10$) |
| ai4i | CNN | 1 | 70–70 | $0.604762 \pm 0.448142$ ($n = 10$) | $0.7375 \pm 0.246151$ ($n = 10$) | $0.7375 \pm 0.246151$ ($n = 10$) | $0.74 \pm 0.350238$ ($n = 10$) |
| ai4i | CNN_LSTM | 0 | 128–128 | $0.785714 \pm 0.189482$ ($n = 10$) | $0.8 \pm 0.344601$ ($n = 10$) | $0.8 \pm 0.344601$ ($n = 10$) | $0.72 \pm 0.388158$ ($n = 10$) |
| ai4i | CNN_LSTM | 1 | 70–70 | $0.395238 \pm 0.638896$ ($n = 10$) | $0.6875 \pm 0.349851$ ($n = 10$) | $0.6875 \pm 0.349851$ ($n = 10$) | $0.71 \pm 0.384274$ ($n = 10$) |
| ai4i | FusionNetLite | 0 | 128–128 | $0.595238 \pm 0.418458$ ($n = 10$) | $0.8 \pm 0.237171$ ($n = 10$) | $0.8 \pm 0.237171$ ($n = 10$) | $0.46 \pm 0.236643$ ($n = 10$) |
| ai4i | FusionNetLite | 1 | 70–70 | $0.795238 \pm 0.11512$ ($n = 10$) | $0.8875 \pm 0.149652$ ($n = 10$) | $0.8875 \pm 0.149652$ ($n = 10$) | $0.69 \pm 0.344642$ ($n = 10$) |
| ai4i | MLP | 0 | 128–128 | $0.595238 \pm 0.457159$ ($n = 10$) | $0.75 \pm 0.204124$ ($n = 10$) | $0.75 \pm 0.204124$ ($n = 10$) | $0.53 \pm 0.349762$ ($n = 10$) |
| ai4i | MLP | 1 | 70–70 | $0.542857 \pm 0.604093$ ($n = 10$) | $0.7375 \pm 0.246151$ ($n = 10$) | $0.7375 \pm 0.246151$ ($n = 10$) | $0.61 \pm 0.357305$ ($n = 10$) |
| metro | BiLSTM | 0 | 128–128 | $0.516429 \pm 0.336828$ ($n = 10$) | $0.78 \pm 0.150882$ ($n = 10$) | $0.817347 \pm 0.164077$ ($n = 7$) | $0.34 \pm 0.217051$ ($n = 10$) |
| metro | BiLSTM | 1 | 128–128 | $0.367857 \pm 0.577343$ ($n = 10$) | $0.66 \pm 0.23402$ ($n = 10$) | $0.657987 \pm 0.241615$ ($n = 10$) | $0.52 \pm 0.385285$ ($n = 10$) |
| metro | CNN | 0 | 128–128 | $0.742143 \pm 0.195672$ ($n = 10$) | $0.88 \pm 0.068853$ ($n = 10$) | $0.88 \pm 0.068853$ ($n = 10$) | $0.56 \pm 0.330656$ ($n = 10$) |
| metro | CNN | 1 | 128–128 | $0.73 \pm 0.242426$ ($n = 10$) | $0.853333 \pm 0.112437$ ($n = 10$) | $0.853333 \pm 0.112437$ ($n = 10$) | $0.53 \pm 0.349762$ ($n = 10$) |
| metro | CNN_LSTM | 0 | 128–128 | $0.332857 \pm 0.524198$ ($n = 10$) | $0.653333 \pm 0.24505$ ($n = 10$) | $0.71685 \pm 0.275579$ ($n = 9$) | $0.39 \pm 0.33483$ ($n = 10$) |
| metro | CNN_LSTM | 1 | 128–128 | $0.307857 \pm 0.466487$ ($n = 10$) | $0.653333 \pm 0.22836$ ($n = 10$) | $0.675705 \pm 0.231574$ ($n = 10$) | $0.25 \pm 0.190029$ ($n = 10$) |
| metro | FusionNetLite | 0 | 128–128 | $0.240714 \pm 0.252$ ($n = 10$) | $0.553333 \pm 0.144188$ ($n = 10$) | $0.633929 \pm 0.183851$ ($n = 8$) | $0.28 \pm 0.311983$ ($n = 10$) |
| metro | FusionNetLite | 1 | 128–128 | $0.275357 \pm 0.501176$ ($n = 10$) | $0.58 \pm 0.220101$ ($n = 10$) | $0.660606 \pm 0.188902$ ($n = 9$) | $0.37 \pm 0.262679$ ($n = 10$) |
| metro | MLP | 0 | 128–128 | $0.527143 \pm 0.485931$ ($n = 10$) | $0.82 \pm 0.160401$ ($n = 10$) | $0.82 \pm 0.160401$ ($n = 10$) | $0.45 \pm 0.32059$ ($n = 10$) |
| metro | MLP | 1 | 128–128 | $0.806429 \pm 0.157636$ ($n = 10$) | $0.853333 \pm 0.12492$ ($n = 10$) | $0.853333 \pm 0.12492$ ($n = 10$) | $0.54 \pm 0.271621$ ($n = 10$) |

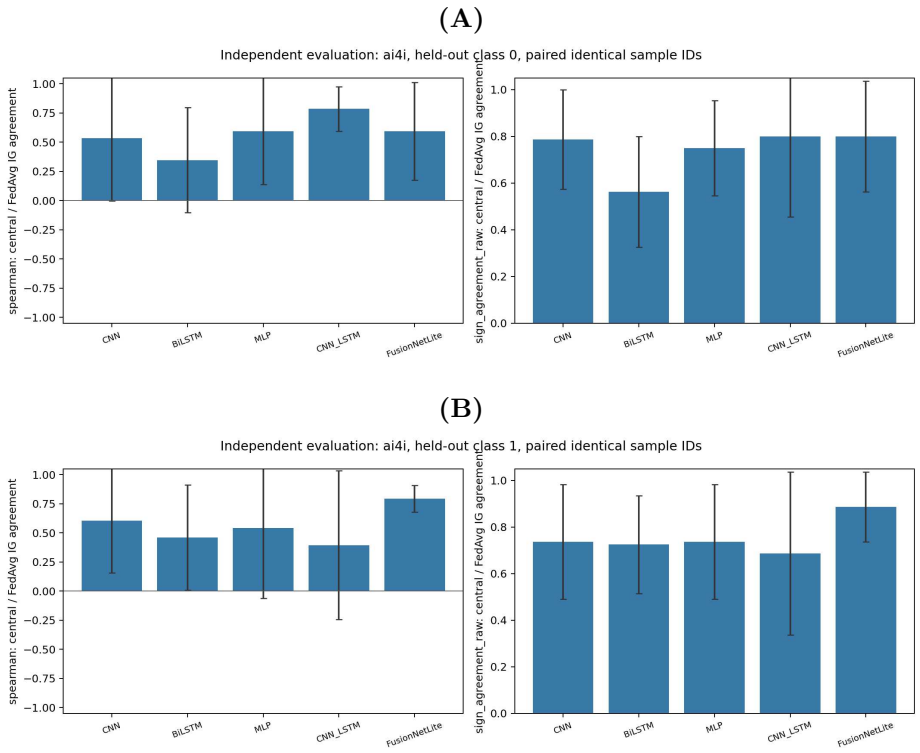



**Fig. 13** AI4I central–FedAvg IG agreement for non-failure (A) and failure (B) observations: Spearman correlation (left) and raw sign agreement (right).

Table 9 reports MetroPT3 FusionNet Lite features with sign-change frequency at least 0.5, separately by class. Among non-failure observations, `DV_eletric` reverses sign in six of ten seed pairs and `TP3` in eight of ten. Signed means, variability and mean changes are retained in scientific notation where needed. The frequencies describe agreement of model attributions rather than the direction or magnitude of a physical sensor effect.

Magnitude filtering raises MetroPT3 FusionNet Lite sign agreement to 0.6339 for class 0 and 0.6606 for class 1, calculated from eight and nine defined seed-level values respectively. Undefined filtered values occur when no features satisfy the recorded magnitude condition and are not replaced with perfect or zero agreement. This distinction matters because a reversal of a near-zero attribution can change a sign statistic without representing a substantial change in model response. The supplementary attribution CSV retains the seed-level agreement values and numbers of features meeting the magnitude condition.

The cross-architecture comparison shows that explanatory agreement is not a direct surrogate for detection quality. On MetroPT3 failures, MLP has mean Spearman correlation $0.8064 \pm 0.1576$, compared with $0.2754 \pm 0.5012$ for FusionNet Lite, despite

**Table 9** MetroPT3 FusionNet Lite features with sign-change frequency at least 0.5.

| Class | MetroPT3 Lite feature | Central IG | FedAvg IG | Mean delta | Raw sign-change rate | Strict reversal rate |
|---|---|---|---|---|---|---|
| 0 | COMP | $-0.0050615 \pm 0.0126622$ | $1.59176 \times 10^{-5} \pm 3.14816 \times 10^{-5}$ | 0.00507742 | 0.5 | 0.5 |
| 0 | DV_eletric | $-0.0111771 \pm 0.0323633$ | $5.76245 \times 10^{-5} \pm 0.000136302$ | 0.0112347 | 0.6 | 0.6 |
| 0 | H1 | $-0.000930517 \pm 0.00191295$ | $2.58227 \times 10^{-5} \pm 0.000102965$ | 0.00095634 | 0.5 | 0.5 |
| 0 | LPS | $-0.0004484 \pm 0.00155799$ | $6.88903 \times 10^{-5} \pm 0.000203496$ | 0.00051729 | 0.7 | 0.7 |
| 0 | MPG | $-0.00139498 \pm 0.00356693$ | $-3.70439 \times 10^{-6} \pm 9.67556 \times 10^{-6}$ | 0.00139128 | 0.5 | 0.5 |
| 0 | Oil_level | $-0.00540166 \pm 0.0188718$ | $1.54805 \times 10^{-5} \pm 4.07607 \times 10^{-5}$ | 0.00541714 | 0.5 | 0.5 |
| 0 | Oil_temperature | $-0.0170807 \pm 0.0515536$ | $5.44632 \times 10^{-6} \pm 3.20992 \times 10^{-5}$ | 0.0170861 | 0.6 | 0.6 |
| 0 | Reservoirs | $-0.0144341 \pm 0.0416739$ | $-5.22043 \times 10^{-7} \pm 9.11477 \times 10^{-6}$ | 0.0144335 | 0.5 | 0.5 |
| 0 | TP3 | $-0.0047159 \pm 0.0136976$ | $-4.80364 \times 10^{-6} \pm 2.34746 \times 10^{-5}$ | 0.00471109 | 0.8 | 0.8 |
| 1 | Caudal_impulses | $-0.00217967 \pm 0.00827558$ | $5.62049 \times 10^{-5} \pm 0.000951194$ | 0.00223587 | 0.6 | 0.6 |
| 1 | H1 | $-0.0375234 \pm 0.0868158$ | $0.0040167 \pm 0.0144531$ | 0.0415401 | 0.6 | 0.6 |
| 1 | MPG | $-0.0137012 \pm 0.0402693$ | $-0.0015803 \pm 0.00339667$ | 0.0121209 | 0.7 | 0.7 |
| 1 | Oil_level | $0.00168563 \pm 0.00301413$ | $-7.69992 \times 10^{-5} \pm 0.000175279$ | −0.00176263 | 0.7 | 0.7 |
| 1 | Pressure_switch | $0.000235886 \pm 0.00201733$ | $-0.000119807 \pm 0.000266838$ | −0.000355693 | 0.6 | 0.6 |
| 1 | TP3 | $0.0196642 \pm 0.0571962$ | $-0.00424773 \pm 0.014848$ | −0.023912 | 0.5 | 0.5 |

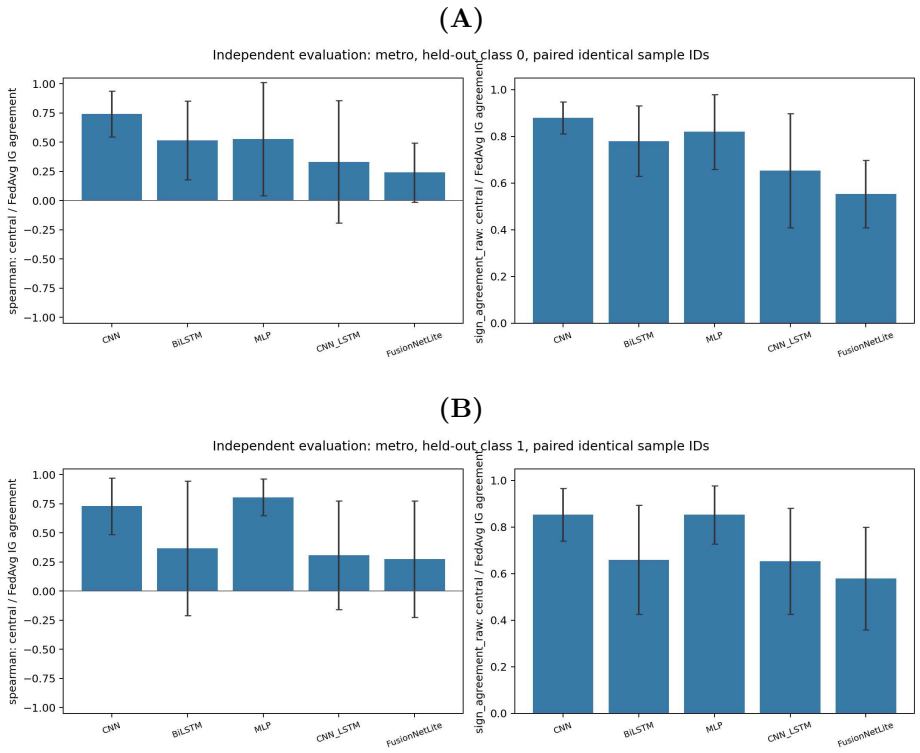



**Fig. 14** MetroPT3 central–FedAvg IG agreement for non-failure (A) and failure (B) observations: Spearman correlation (left) and raw sign agreement (right).

the MLP's weak predictive performance. BiLSTM has $0.3679 \pm 0.5773$, demonstrating that a recurrent baseline is not uniformly stable either. These findings apply to the selected IG baseline, paired samples and model checkpoints; they do not establish that one attribution method captures every aspect of model reasoning. This separates explanation agreement from predictive utility, consistent with the need for quantitative attribution comparison identified in Sect. 2 and the paired design in Sect. 3.8.

## 4.9 Decision policies and imbalance-sensitive training

Tables 10 and 11 compare four decision policies at fixed best checkpoints for FusionNet Lite and MLP. Tables 12 and 13 report their corresponding test confusion counts. Entries report means and sample SD across ten seeds; confusion counts are averaged across seeds and do not form a pooled confusion matrix. For MetroPT3 FedAvg FusionNet Lite, mean test $F_2$ is $1.5404 \times 10^{-4}$ at a fixed threshold of 0.50, $0.0423 \pm 0.0995$ under the original validation-$F_1$ grid, and $0.1566 \pm 0.2044$ under expanded validation-$F_2$ selection. The corresponding mean recalls are 0.0001233, 0.0421 and 0.1887. Expanded validation selection chooses a mean threshold of 0.0029, below the original grid's 0.05 lower bound; this is an average of seed-specific selections, not a fixed

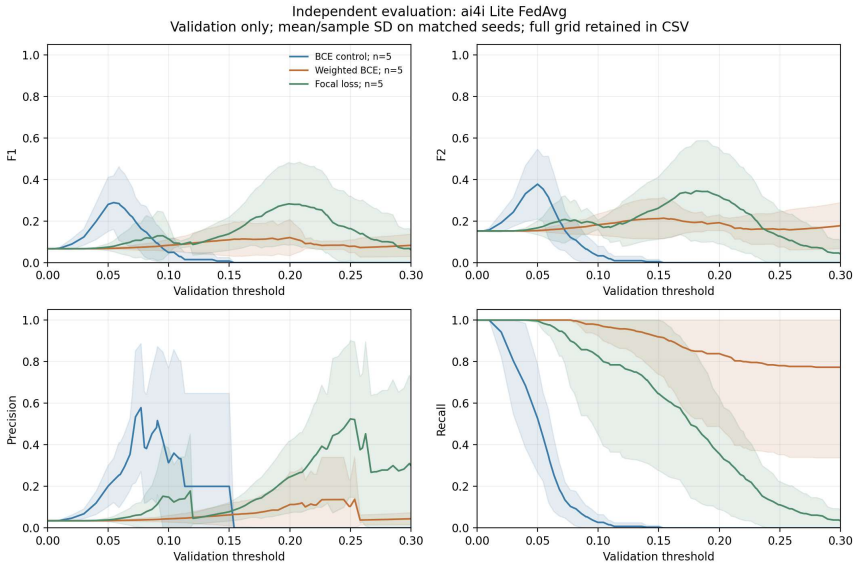



**Fig. 15** AI4I validation threshold sweeps under three losses: $F_1$, $F_2$, precision and recall, in reading order.

deployment threshold. These results identify a consequential operating-point limitation, although the large $F_2$ dispersion shows that threshold expansion does not reliably recover failures for every seed. For centralized MetroPT3 FusionNet Lite, expanded validation-$F_2$ selection also does not exceed expanded validation-$F_1$ test $F_2$ on average, illustrating the distinction between optimizing a validation objective and guaranteeing its transfer to later observations.

Loss sensitivity provides a complementary intervention within training. Across the matched five seeds (5–9), MetroPT3 FedAvg FusionNet Lite has mean test $F_2$ $0.0835 \pm 0.1217$ under unweighted BCE, $0.2620 \pm 0.2433$ under weighted BCE and $0.3158 \pm 0.1828$ under focal loss, all evaluated with expanded validation-$F_2$ thresholds. The loss conditions therefore yield different descriptive means under the same validation-$F_2$ selection policy, while the remaining variance limits a general deployment claim. Figures 15 and 16 show validation-only threshold sweeps for these five-seed loss comparisons: unweighted BCE, weighted BCE and focal loss. Lines and shading represent means and sample SD. These plots display validation $F_1$, $F_2$, precision and recall in a $2\times2$ layout; they are not the ten-seed test operating-point comparison. The supplementary predictive CSV includes the matched test loss and decision-policy conditions.

**Table 10** Lite/MLP operating points at best validation checkpoints, ten seeds. Thresholds are selected using validation only; fixed 0.5 is diagnostic. Means and sample SD across seeds; all per-seed/all-model counts remain in the canonical CSV. Part 1 of 2.

| Dataset | Model | Mode | Policy | Threshold | Precision | Recall | $F_2$ |
|---|---|---|---|---|---|---|---|
| ai4i | FusionNetLite | central | expanded_f1 | $0.166231 \pm 0.144441$ | $0.20179 \pm 0.119465$ | $0.53 \pm 0.243412$ | $0.322911 \pm 0.129397$ |
| ai4i | FusionNetLite | central | expanded_f2 | $0.149402 \pm 0.151507$ | $0.172201 \pm 0.113865$ | $0.651429 \pm 0.15825$ | $0.361708 \pm 0.1279$ |
| ai4i | FusionNetLite | central | fixed_0_5 | $0.5 \pm 0$ | $0.102207 \pm 0.315529$ | $0.04 \pm 0.116798$ | $0.0124624 \pm 0.0291632$ |
| ai4i | FusionNetLite | central | original_f1 | $0.166231 \pm 0.144441$ | $0.20179 \pm 0.119465$ | $0.53 \pm 0.243412$ | $0.322911 \pm 0.129397$ |
| ai4i | FusionNetLite | fedavg | expanded_f1 | $0.0561407 \pm 0.0172186$ | $0.234016 \pm 0.110008$ | $0.404286 \pm 0.177415$ | $0.319694 \pm 0.116995$ |
| ai4i | FusionNetLite | fedavg | expanded_f2 | $0.0423568 \pm 0.0123701$ | $0.139624 \pm 0.0609799$ | $0.61 \pm 0.129721$ | $0.350832 \pm 0.0908306$ |
| ai4i | FusionNetLite | fedavg | fixed_0_5 | $0.5 \pm 0$ | $0 \pm 0$ | $0 \pm 0$ | $0 \pm 0$ |
| ai4i | FusionNetLite | fedavg | original_f1 | $0.060402 \pm 0.0116871$ | $0.311365 \pm 0.270745$ | $0.284286 \pm 0.195464$ | $0.255478 \pm 0.158385$ |
| ai4i | MLP | central | expanded_f1 | $0.336356 \pm 0.201929$ | $0.235257 \pm 0.216382$ | $0.555714 \pm 0.272758$ | $0.337674 \pm 0.237761$ |
| ai4i | MLP | central | expanded_f2 | $0.307793 \pm 0.21081$ | $0.193869 \pm 0.169985$ | $0.64 \pm 0.225656$ | $0.35175 \pm 0.207476$ |
| ai4i | MLP | central | fixed_0_5 | $0.5 \pm 0$ | $0.363522 \pm 0.41817$ | $0.337143 \pm 0.231842$ | $0.216102 \pm 0.141159$ |
| ai4i | MLP | central | original_f1 | $0.341256 \pm 0.193297$ | $0.235257 \pm 0.216382$ | $0.555714 \pm 0.272758$ | $0.337674 \pm 0.237761$ |
| ai4i | MLP | fedavg | expanded_f1 | $0.203769 \pm 0.187905$ | $0.247454 \pm 0.0979358$ | $0.281429 \pm 0.114296$ | $0.263495 \pm 0.0886527$ |
| ai4i | MLP | fedavg | expanded_f2 | $0.16497 \pm 0.202348$ | $0.138104 \pm 0.0501267$ | $0.484286 \pm 0.130488$ | $0.313113 \pm 0.0750099$ |
| ai4i | MLP | fedavg | fixed_0_5 | $0.5 \pm 0$ | $0.00897516 \pm 0.0193445$ | $0.102857 \pm 0.248424$ | $0.0298755 \pm 0.0630048$ |
| ai4i | MLP | fedavg | original_f1 | $0.203769 \pm 0.187905$ | $0.247454 \pm 0.0979358$ | $0.281429 \pm 0.114296$ | $0.263495 \pm 0.0886527$ |
| metro | FusionNetLite | central | expanded_f1 | $0.0018 \pm 0.00168655$ | $0.269061 \pm 0.100484$ | $0.422873 \pm 0.168514$ | $0.357274 \pm 0.104195$ |
| metro | FusionNetLite | central | expanded_f2 | $0.0014 \pm 0.00126491$ | $0.248141 \pm 0.105676$ | $0.451171 \pm 0.193534$ | $0.353707 \pm 0.102586$ |
| metro | FusionNetLite | central | fixed_0_5 | $0.5 \pm 0$ | $0.0924051 \pm 0.29221$ | $0.00450062 \pm 0.0142322$ | $0.00555809 \pm 0.0175762$ |
| metro | FusionNetLite | central | original_f1 | $0.05 \pm 0$ | $0.463177 \pm 0.342094$ | $0.0821825 \pm 0.0918125$ | $0.0966051 \pm 0.106609$ |

**Table 11** FusionNet Lite and MLP test operating points (part 2 of 2).

| Dataset | Model | Mode | Policy | Threshold | Precision | Recall | $F_2$ |
|---|---|---|---|---|---|---|---|
| metro | FusionNetLite | fedavg | expanded_f1 | $0.0039 \pm 0.00917061$ | $0.2636 \pm 0.299261$ | $0.169975 \pm 0.236844$ | $0.155685 \pm 0.203426$ |
| metro | FusionNetLite | fedavg | expanded_f2 | $0.0029 \pm 0.00600833$ | $0.259464 \pm 0.301115$ | $0.188718 \pm 0.267403$ | $0.156635 \pm 0.204369$ |
| metro | FusionNetLite | fedavg | fixed_0_5 | $0.5 \pm 0$ | $0.05 \pm 0.158114$ | $0.000123305 \pm 0.000389923$ | $0.000154036 \pm 0.000487104$ |
| metro | FusionNetLite | fedavg | original_f1 | $0.05 \pm 0$ | $0.180443 \pm 0.31591$ | $0.0421085 \pm 0.104073$ | $0.0422504 \pm 0.0995371$ |
| metro | MLP | central | expanded_f1 | $0.0037 \pm 0.00596378$ | $0.0002457 \pm 0.000776972$ | $6.16523 \times 10^{-5} \pm 0.000194962$ | $7.25163 \times 10^{-5} \pm 0.000229317$ |
| metro | MLP | central | expanded_f2 | $0.0022 \pm 0.00193218$ | $0.0002457 \pm 0.000776972$ | $6.16523 \times 10^{-5} \pm 0.000194962$ | $7.25163 \times 10^{-5} \pm 0.000229317$ |
| metro | MLP | central | fixed_0_5 | $0.5 \pm 0$ | $0 \pm 0$ | $0 \pm 0$ | $0 \pm 0$ |
| metro | MLP | central | original_f1 | $0.05 \pm 0$ | $0 \pm 0$ | $0 \pm 0$ | $0 \pm 0$ |
| metro | MLP | fedavg | expanded_f1 | $0.0022 \pm 0.00193218$ | $0.00700662 \pm 0.0112848$ | $0.000246609 \pm 0.000431076$ | $0.000305253 \pm 0.000532813$ |
| metro | MLP | fedavg | expanded_f2 | $0.001 \pm 2.2857 \times 10^{-19}$ | $0.00473022 \pm 0.00970398$ | $0.000308261 \pm 0.000523944$ | $0.000347717 \pm 0.000587683$ |
| metro | MLP | fedavg | fixed_0_5 | $0.5 \pm 0$ | $0 \pm 0$ | $0 \pm 0$ | $0 \pm 0$ |
| metro | MLP | fedavg | original_f1 | $0.05 \pm 0$ | $0 \pm 0$ | $0 \pm 0$ | $0 \pm 0$ |

**Table 12** Matched Lite/MLP test confusion counts for the same operating-point policies. Entries are mean counts and sample SD across ten seeds, not one pooled confusion matrix. No test labels are used to choose thresholds. Part 1 of 2.

| Dataset | Model | Mode | Policy | TN | FP | FN | TP |
|---|---|---|---|---|---|---|---|
| ai4i | FusionNetLite | central | expanded_f1 | $1533 \pm 632.82$ | $398 \pm 632.82$ | $32.9 \pm 17.0388$ | $37.1 \pm 17.0388$ |
| ai4i | FusionNetLite | central | expanded_f2 | $1467.6 \pm 600.844$ | $463.4 \pm 600.844$ | $24.4 \pm 11.0775$ | $45.6 \pm 11.0775$ |
| ai4i | FusionNetLite | central | fixed_0_5 | $1815.8 \pm 364.294$ | $115.2 \pm 364.294$ | $67.2 \pm 8.17585$ | $2.8 \pm 8.17585$ |
| ai4i | FusionNetLite | central | original_f1 | $1533 \pm 632.82$ | $398 \pm 632.82$ | $32.9 \pm 17.0388$ | $37.1 \pm 17.0388$ |
| ai4i | FusionNetLite | fedavg | expanded_f1 | $1798.2 \pm 119.005$ | $132.8 \pm 119.005$ | $41.7 \pm 12.4191$ | $28.3 \pm 12.4191$ |
| ai4i | FusionNetLite | fedavg | expanded_f2 | $1625.2 \pm 145.882$ | $305.8 \pm 145.882$ | $27.3 \pm 9.0805$ | $42.7 \pm 9.0805$ |
| ai4i | FusionNetLite | fedavg | fixed_0_5 | $1931 \pm 0$ | $0 \pm 0$ | $70 \pm 0$ | $0 \pm 0$ |
| ai4i | FusionNetLite | fedavg | original_f1 | $1864.6 \pm 72.5277$ | $66.4 \pm 72.5277$ | $50.1 \pm 13.6825$ | $19.9 \pm 13.6825$ |
| ai4i | MLP | central | expanded_f1 | $1447.8 \pm 685.438$ | $483.2 \pm 685.438$ | $31.1 \pm 19.093$ | $38.9 \pm 19.093$ |
| ai4i | MLP | central | expanded_f2 | $1348.6 \pm 726.9$ | $582.4 \pm 726.9$ | $25.2 \pm 15.7959$ | $44.8 \pm 15.7959$ |
| ai4i | MLP | central | fixed_0_5 | $1599.9 \pm 506.051$ | $331.1 \pm 506.051$ | $46.4 \pm 16.2289$ | $23.6 \pm 16.2289$ |
| ai4i | MLP | central | original_f1 | $1447.8 \pm 685.438$ | $483.2 \pm 685.438$ | $31.1 \pm 19.093$ | $38.9 \pm 19.093$ |
| ai4i | MLP | fedavg | expanded_f1 | $1862.2 \pm 41.9995$ | $68.8 \pm 41.9995$ | $50.3 \pm 8.00069$ | $19.7 \pm 8.00069$ |
| ai4i | MLP | fedavg | expanded_f2 | $1700.1 \pm 85.9282$ | $230.9 \pm 85.9282$ | $36.1 \pm 9.13418$ | $33.9 \pm 9.13418$ |
| ai4i | MLP | fedavg | fixed_0_5 | $1755.9 \pm 452.866$ | $175.1 \pm 452.866$ | $62.8 \pm 17.3897$ | $7.2 \pm 17.3897$ |
| ai4i | MLP | fedavg | original_f1 | $1862.2 \pm 41.9995$ | $68.8 \pm 41.9995$ | $50.3 \pm 8.00069$ | $19.7 \pm 8.00069$ |
| metro | FusionNetLite | central | expanded_f1 | $375355 \pm 1853.87$ | $2260.6 \pm 1853.87$ | $936.1 \pm 273.329$ | $685.9 \pm 273.329$ |
| metro | FusionNetLite | central | expanded_f2 | $374716 \pm 2505.41$ | $2899.9 \pm 2505.41$ | $890.2 \pm 313.913$ | $731.8 \pm 313.913$ |
| metro | FusionNetLite | central | fixed_0_5 | $377609 \pm 6.9602$ | $7 \pm 6.9602$ | $1614.7 \pm 23.0846$ | $7.3 \pm 23.0846$ |
| metro | FusionNetLite | central | original_f1 | $377517 \pm 102.163$ | $99 \pm 102.163$ | $1488.7 \pm 148.92$ | $133.3 \pm 148.92$ |

**Table 13** FusionNet Lite and MLP test confusion counts (part 2 of 2).

| Dataset | Model | Mode | Policy | TN | FP | FN | TP |
|---|---|---|---|---|---|---|---|
| metro | FusionNetLite | fedavg | expanded_f1 | $376930 \pm 1228.32$ | $686.4 \pm 1228.32$ | $1346.3 \pm 384.16$ | $275.7 \pm 384.16$ |
| metro | FusionNetLite | fedavg | expanded_f2 | $376548 \pm 2377.06$ | $1068.3 \pm 2377.06$ | $1315.9 \pm 433.727$ | $306.1 \pm 433.727$ |
| metro | FusionNetLite | fedavg | fixed_0_5 | $377616 \pm 1.08012$ | $0.5 \pm 1.08012$ | $1621.8 \pm 0.632456$ | $0.2 \pm 0.632456$ |
| metro | FusionNetLite | fedavg | original_f1 | $377450 \pm 505.797$ | $165.7 \pm 505.797$ | $1553.7 \pm 168.807$ | $68.3 \pm 168.807$ |
| metro | MLP | central | expanded_f1 | $377540 \pm 123.17$ | $76.1 \pm 123.17$ | $1621.9 \pm 0.316228$ | $0.1 \pm 0.316228$ |
| metro | MLP | central | expanded_f2 | $377508 \pm 153.476$ | $108.1 \pm 153.476$ | $1621.9 \pm 0.316228$ | $0.1 \pm 0.316228$ |
| metro | MLP | central | fixed_0_5 | $377608 \pm 24.4595$ | $8.4 \pm 24.4595$ | $1622 \pm 0$ | $0 \pm 0$ |
| metro | MLP | central | original_f1 | $377588 \pm 65.896$ | $27.5 \pm 65.896$ | $1622 \pm 0$ | $0 \pm 0$ |
| metro | MLP | fedavg | expanded_f1 | $377558 \pm 83.3521$ | $57.7 \pm 83.3521$ | $1621.6 \pm 0.699206$ | $0.4 \pm 0.699206$ |
| metro | MLP | fedavg | expanded_f2 | $377298 \pm 601.878$ | $318.1 \pm 601.878$ | $1621.5 \pm 0.849837$ | $0.5 \pm 0.849837$ |
| metro | MLP | fedavg | fixed_0_5 | $377613 \pm 8.50882$ | $2.8 \pm 8.50882$ | $1622 \pm 0$ | $0 \pm 0$ |
| metro | MLP | fedavg | original_f1 | $377593 \pm 63.9528$ | $23.2 \pm 63.9528$ | $1622 \pm 0$ | $0 \pm 0$ |

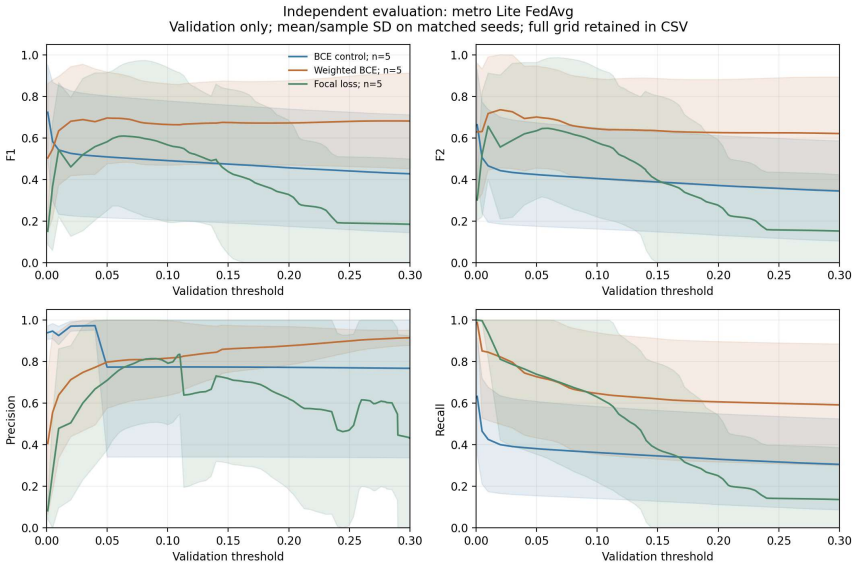



**Fig. 16** MetroPT3 validation threshold sweeps under three losses: $F_1$, $F_2$, precision and recall, in reading order.

## 4.10 Component and feature-order sensitivity

Table 14 compares the full architecture with five component or convolution variants over the same seeds (5–9). The component effects are conditional rather than uniformly beneficial. Removing SE attention increases centralized AI4I AUPRC from the matched full-model mean of 0.2185 to 0.4264 and $F_2$ from 0.3409 to 0.4980. In contrast, removing the stem residual decreases centralized MetroPT3 $F_2$ from 0.3717 to 0.0738. Standard convolution attains MetroPT3 $F_2$ of 0.3450 centrally and 0.1920 under FedAvg, compared with 0.3717 and 0.0835 for the matched full model. These comparisons support assessing each component's contribution under its operating condition; inclusion alone does not demonstrate a universal benefit.

Feature-order perturbations also change the outcome. For centralized MetroPT3 FusionNet Lite, the fixed permutation yields $F_2$ 0.1132 compared with 0.3717 under canonical order and 0.3646 under reverse order. Under FedAvg, the corresponding means are 0.2807, 0.0835 and 0.2378. These convolutions mix neighbouring feature positions, not successive temporal observations, making order dependence a limitation of the chosen tabular representation. These matched conditions are retained in `full_canonical_metrics.csv`; they do not establish a uniquely optimal ordering. The observed sensitivity is consistent with the feature-axis representation in Sect. 3.1, rather than a temporal forecasting interpretation.

**Table 14** FusionNet Lite component ablations and measured resources across five matched seeds.

| Dataset | Mode | Variant (full is control) | Seeds | Parameters | AUPRC | $F_2$ (val $F_2$) | Latency (ms/sample) |
|---|---|---|---|---|---|---|---|
| ai4i | central | full | 5 | 1420 | $0.218536 \pm 0.202935$ | $0.34091 \pm 0.17306$ | $0.164418 \pm 0.0103142$ |
| ai4i | central | no_mixer | 5 | 1020 | $0.232494 \pm 0.110095$ | $0.345621 \pm 0.0670356$ | $0.163053 \pm 0.00810181$ |
| ai4i | central | no_mixer_residual | 5 | 1420 | $0.385724 \pm 0.191461$ | $0.453839 \pm 0.136795$ | $0.165612 \pm 0.0114038$ |
| ai4i | central | no_se | 5 | 1140 | $0.426375 \pm 0.0869725$ | $0.497997 \pm 0.0616121$ | $0.161721 \pm 0.0112891$ |
| ai4i | central | no_stem_residual | 5 | 1388 | $0.125206 \pm 0.0852603$ | $0.291416 \pm 0.091104$ | $0.161684 \pm 0.00843224$ |
| ai4i | central | standard_conv | 5 | 1913 | $0.351104 \pm 0.247398$ | $0.431352 \pm 0.19211$ | $0.159748 \pm 0.00532319$ |
| ai4i | fedavg | full | 5 | 1420 | $0.265448 \pm 0.146061$ | $0.378941 \pm 0.103306$ | $0.155203 \pm 0.00763058$ |
| ai4i | fedavg | no_mixer | 5 | 1020 | $0.0942573 \pm 0.063423$ | $0.223445 \pm 0.058124$ | $0.159721 \pm 0.00506618$ |
| ai4i | fedavg | no_mixer_residual | 5 | 1420 | $0.282607 \pm 0.0988688$ | $0.365894 \pm 0.119701$ | $0.174189 \pm 0.00745876$ |
| ai4i | fedavg | no_se | 5 | 1140 | $0.29914 \pm 0.124639$ | $0.376417 \pm 0.135464$ | $0.165759 \pm 0.0121304$ |
| ai4i | fedavg | no_stem_residual | 5 | 1388 | $0.178573 \pm 0.0833773$ | $0.285755 \pm 0.0767351$ | $0.163591 \pm 0.00813658$ |
| ai4i | fedavg | standard_conv | 5 | 1913 | $0.258116 \pm 0.0940379$ | $0.376751 \pm 0.0612784$ | $0.164198 \pm 0.0138395$ |
| metro | central | full | 5 | 1420 | $0.328471 \pm 0.155859$ | $0.371652 \pm 0.0782316$ | $0.163588 \pm 0.00728313$ |
| metro | central | no_mixer | 5 | 1020 | $0.137213 \pm 0.140589$ | $0.19261 \pm 0.224981$ | $0.16207 \pm 0.0128281$ |
| metro | central | no_mixer_residual | 5 | 1420 | $0.340406 \pm 0.186124$ | $0.35024 \pm 0.249407$ | $0.171038 \pm 0.00630809$ |
| metro | central | no_se | 5 | 1140 | $0.253488 \pm 0.235476$ | $0.218457 \pm 0.246923$ | $0.171143 \pm 0.0113119$ |
| metro | central | no_stem_residual | 5 | 1388 | $0.134196 \pm 0.203252$ | $0.0738117 \pm 0.147768$ | $0.168374 \pm 0.00922348$ |
| metro | central | standard_conv | 5 | 1913 | $0.294653 \pm 0.162041$ | $0.345031 \pm 0.18842$ | $0.167597 \pm 0.012012$ |
| metro | fedavg | full | 5 | 1420 | $0.1862 \pm 0.231032$ | $0.0835351 \pm 0.121689$ | $0.15932 \pm 0.00524088$ |
| metro | fedavg | no_mixer | 5 | 1020 | $0.172196 \pm 0.122041$ | $0.19674 \pm 0.184456$ | $0.163988 \pm 0.0127044$ |
| metro | fedavg | no_mixer_residual | 5 | 1420 | $0.261239 \pm 0.183115$ | $0.235089 \pm 0.241022$ | $0.162071 \pm 0.00427199$ |
| metro | fedavg | no_se | 5 | 1140 | $0.333725 \pm 0.157551$ | $0.124562 \pm 0.164655$ | $0.159315 \pm 0.00832493$ |
| metro | fedavg | no_stem_residual | 5 | 1388 | $0.129548 \pm 0.167829$ | $0.0505524 \pm 0.112611$ | $0.165863 \pm 0.0115382$ |
| metro | fedavg | standard_conv | 5 | 1913 | $0.371567 \pm 0.107546$ | $0.192031 \pm 0.196904$ | $0.163895 \pm 0.0110477$ |

## 4.11 Federation size, heterogeneity and participation

Following Sect. 3.7, the sensitivity analysis includes three and ten simulated clients, Dirichlet $\alpha \in \{0.1, 0.5, 10\}$ and separate quantity-allocation controls. With equal target client fractions, MetroPT3 FusionNet Lite $F_2$ ranges from 0.0794 to 0.2747 across the three-client $\alpha$ conditions and from 0.1197 to 0.2157 across the ten-client conditions. The ordering is not monotonic in $\alpha$ across client counts, and each setting uses only three matched seeds. Client count or label concentration therefore does not provide a single explanation for the observed detection variability.

At ten clients and $\alpha = 0.5$, MetroPT3 mean $F_2$ changes from 0.2157 under full participation to 0.1889 at 50% participation, and to 0.1094 when independent 20% dropout is added. The AI4I means do not follow the same ordering. These results characterize the fixed 20-round simulation budget; they do not measure asynchronous industrial connectivity or behavior of 50-node fleets. The supplementary client-heterogeneity CSV records the realized client distributions; the predictive CSV retains the participation conditions.

## 4.12 Model-specific learning-rate sensitivity

The common learning rate provides a controlled core comparison but does not guarantee that each architecture is well tuned, as distinguished in Sect. 3.2. In the three-seed sensitivity study, learning rates are selected using mean best validation AUROC, without reference to test metrics. For example, the selected centralized AI4I CNN at learning rate $3 \times 10^{-3}$ obtains AUROC $0.9340 \pm 0.0125$, AUPRC $0.5647 \pm 0.0223$ and $F_2$ $0.5578 \pm 0.0225$. Model ranking is consequently sensitive to optimization configuration. These three-seed estimates are reported separately from the ten-seed core results, with candidate-rate conditions retained in the supplementary predictive CSV.

## 4.13 Fixed-checkpoint gradient diagnostics

Training-only diagnostics examine deterministic client batches at the same best global checkpoint without updating the model, following Sect. 3.7. Across the five diagnostic seeds, the 15 pairwise client-gradient cosine values for FusionNet Lite average $-0.2735$ on AI4I and $-0.1754$ on MetroPT3, with both strongly opposing and closely aligned pairs in each dataset. MLP averages are 0.1346 and 0.1129, although its pairwise ranges also include opposition. Matched batches include zero-positive clients. These observations establish that client-gradient disagreement and scarce failure examples occur in the evaluated settings; they do not demonstrate that gradient starvation or a particular sensor correlation causes an attribution reversal. These training-only checks remain descriptive diagnostics.

## 4.14 Overall empirical interpretation

The results support a conditional performance–complexity trade-off. FusionNet Lite has the smallest parameter count and fixed-batch EEI among the evaluated models, but parameter count does not guarantee the lowest latency or smallest serialized file. Its AI4I FedAvg means are competitive under the core configuration, whereas MetroPT3

exposes the difference between strong ranking discrimination and reliable failure detection. Threshold expansion and alternative losses improve several descriptive detection estimates, with substantial variation remaining across seeds.

The controlled studies also qualify the architectural interpretation. Component removal, feature ordering, client allocation and learning rate can change model rankings; no single component or setting explains every outcome. Class-specific IG agreement adds evidence about explanatory consistency, but neither high agreement nor client-gradient opposition establishes predictive quality or a causal mechanism. These findings answer the comparison and explainability questions raised in Sect. 2 within the defined benchmark, while leaving broader industrial generalization and deployment claims to separate validation.

# 5 Discussion

FusionNet Lite addresses the resource-conscious design question developed in Sect. 2, while the matched comparisons establish a conditional advantage. Among the five core architectures, its 1,420 trainable parameters give it the smallest parameter footprint, and it has the lowest measured fixed-batch EEI in both training modes on both datasets (Table 7). These advantages do not imply the lowest latency, smallest serialized checkpoint or strongest failure detection. The distinction matters for lightweight PdM: parameter reduction is a design property, whereas computational cost and predictive utility require separate measurements under the intended operating conditions.

The MetroPT3 results connect the literature review's concerns about temporal evaluation and class imbalance. High held-out AUROC coexists with substantially weaker and variable AUPRC and $F_2$, so ranking discrimination alone does not establish useful failure detection. The threshold comparison distinguishes the fixed 0.50 diagnostic from validation-selected policies whose original grid began at 0.05. Expanding that grid permits lower operating points, but the mean selected threshold of 0.0029 for federated FusionNet Lite is a summary across seeds, not a universal setting. Alternative losses also change several descriptive detection estimates under matched conditions. Neither intervention guarantees reliable transfer from validation to the later test period, and the validation–test gap does not identify its cause (Fig. 7).

The controlled studies further qualify an architecture-only interpretation of federated performance. Component removal, feature order, client allocation, participation and learning rate change the observed outcomes, sometimes favouring a variant or comparator over the full model. Measured heterogeneity and zero-positive clients establish relevant challenges without isolating one mechanism responsible for every performance change. Four of the 16 planned contrasts survive Holm adjustment, supporting specific differences under the evaluated settings rather than uniform superiority. This evidence gives a bounded answer to the fair-comparison and heterogeneity questions in Sect. 2: model design, optimization and decision policy must be interpreted together.

The attribution results extend this interpretation to explanations. Class-specific central–FedAvg agreement differs across models and datasets, and high agreement can occur for a weak detector (Table 8). For MetroPT3 FusionNet Lite, the mean rank

correlations are positive but variable, indicating uneven agreement across seeds. Magnitude filtering also shows why near-zero sign reversals require caution. Client-gradient disagreement provides a descriptive diagnostic, but does not show that gradient starvation causes the attribution changes. These findings connect directly to the literature review's distinction between predictive and explanatory consistency: reliable maintenance support requires evidence about detection, resources and explanation behaviour, without treating one as a substitute for another.

# 6 Limitations

The principal boundary is external validity. AI4I is synthetic, and its target here is the logical OR of five failure indicators, which should not be assumed equivalent to every study using the original failure field. MetroPT3 represents one compressor recording evaluated through a single fixed-row chronological holdout. Although the recovered timestamp rule reproduces the accepted labels, the authoritative annotation document and historical producing version remain unverified. Chronological separation therefore supports a test on later observations within this recording; it does not establish transfer across independent machines, factories or failure episodes. The single-observation representation also provides no specified forecasting horizon or demonstrated warning lead time. These limits qualify the temporal-generalization and industrial-applicability concerns identified in Sect. 2.

Experimental coverage remains bounded despite the broader campaign. Three- and ten-client simulations, varying label allocation and partial participation extend the core comparison, but do not reproduce physical cross-factory networks, asynchronous updates or heterogeneous devices. FedAvg is the only aggregation method evaluated, with a fixed 20-round budget and three local epochs. The bounded learning-rate search does not establish optimal tuning for every architecture, and the smaller diagnostic and sensitivity seed sets provide less precise estimates than the core comparison. Holm adjustment covers the 16 planned contrasts rather than every exploratory analysis. Repeated seeds characterize variation within the stated protocols, while MetroPT3 retains a shared test interval; they are not independent industrial replications.

The measured quantities also have distinct operational limits. GPU-host timings include API overhead, and EEI combines parameter count with amortized latency; neither establishes device-level energy use, network traffic or compliance with real-time deadlines. Keeping raw data local provides no demonstrated protection against information leakage from exchanged updates, because secure aggregation and differential privacy are not implemented. IG covers all five model families, but remains one attribution method with a zero reference input and finite class-specific samples. Undefined comparisons and small attribution magnitudes constrain interpretation, while neither agreement statistics nor fixed-checkpoint gradient diagnostics establish causal sensor effects or usefulness to maintenance personnel. Reproducibility of this benchmark consequently remains separate from validation of a deployed maintenance system.

# 7 Future Work

The next stage should validate the full decision pipeline on independent machines and additional chronological periods before prospective deployment. This should begin with verified annotation lineage and explicit definitions of failure episodes and warning horizons, followed by rolling temporal evaluation with validation-only threshold selection. Operational assessment should report missed failures, false alarms and warning lead time alongside AUPRC and $F_2$. Deployment on the intended edge hardware should then measure single-observation and batch latency, memory, power and transmitted update volume under realistic connectivity. Operator review would determine whether alerts and explanations support maintenance decisions; secure aggregation and, where appropriate, differentially private training should be evaluated together with their utility and computation costs.

Algorithmic extensions should be tested as controlled alternatives to synchronous FedAvg. FedProx, SCAFFOLD and adaptive server optimizers such as FedAdam provide distinct comparisons for proximal regularization, client-drift correction and server optimization; personalized models should be assessed separately when local objectives differ. Each comparison should retain matched partitions, tuning budgets and communication accounting across label, feature and availability shifts. Representation studies should test feature-order sensitivity and introduce explicit temporal windows before considering temporal backbones. Explanation evaluation should also vary IG reference inputs and include complementary attribution methods, paired class-specific samples and targeted perturbation tests. These studies would test whether proposed mechanisms improve detection, resource use or explanation reliability, without assuming that gains in one dimension transfer to the others.

# Acknowledgments

The authors acknowledge Swinburne University of Technology for institutional and computational support, and the creators of AI4I 2020 and MetroPT3 for making their datasets available for research [18, 22].

This work was performed on the OzSTAR national facility at Swinburne University of Technology. The OzSTAR program receives funding in part from the Astronomy National Collaborative Research Infrastructure Strategy (NCRIS) allocation provided by the Australian Government, and from the Victorian Higher Education State Investment Fund (VHESIF) provided by the Victorian Government.

### *Author Contributions*

**Aman Sharma** (ORCID: 0000-0002-8058-1276) designed the study, implemented the models, conducted all experiments, performed the analysis, and led the manuscript writing. **Kwan Yong Sim** (ORCID: 0000-0002-7291-0922) provided supervision, methodological feedback, and manuscript review. **Sivachandran Chandrasekaran** (ORCID: 0000-0003-2871-880X) contributed supervision, domain expertise, and critical revisions. All authors read and approved the final manuscript.

***Funding***

This research is supported by Swinburne University of Technology, Hawthorn campus, through a PhD scholarship and the provision of laboratory equipment and computational resources. Open-access publication support is enabled through Swinburne's institutional partnership with the journal.

***Data and Code Availability***

The AI4I 2020 and MetroPT3 datasets used in this study are publicly available through their respective source repositories [18, 22]. The complete experimental implementation, environment specification, deterministic runner, preprocessing and training pipeline, evaluation procedures, reference result artifacts, and consistency-audit workflow are available through the accompanying Code Ocean capsule (version 2.0) at https://doi.org/10.24433/CO.2155937.v2. The capsule does not redistribute the raw datasets; instead, it provides dataset placeholders and mounting instructions so authorized copies of AI4I 2020 and MetroPT3 can be attached before executing the full reproducible workflow.

## Supplementary Material

Four standalone CSV files accompany this article. `full_canonical_metrics.csv` provides seed-level predictive metrics, thresholds and confusion counts, with identifiers for the experimental conditions and checkpoints. `full_ig_agreement_seeds.csv` provides class-specific, seed-level central–FedAvg IG agreement metrics and sign-change counts. `significance.csv` reports paired-difference summaries and intervals, together with raw and Holm-adjusted Wilcoxon probabilities for the declared comparisons. `full_client_heterogeneity.csv` records client sample sizes, positive counts, class prevalence and total-variation distances under the evaluated allocation conditions.

## Declarations

***AI-assisted tools***

OpenAI Codex (GPT-6.1 Sol Ultra) supported code interpretation, debugging, verification and extraction of the experimental implementation, together with source-backed manuscript editing and scripts for numerical tables and figures. We used its coding assistance to check the workflow before computationally intensive runs on shared resources. The authors retain responsibility for the methods, source verification, numerical results, interpretation and final text. Grammarly Premium was used for sentence-level copyediting, including grammar, wording, and readability.

***Competing Interests***

The authors declare that they have no competing interests.

***Consent for Publication***

All authors have read and agreed to the publishing of the manuscript.

### *Research Involving Human Participants and/or Animals*

This research does not involve human participants or animals.

### *Informed Consent*

Not applicable, as the study does not involve human participants.

## References


[1] Leng, J., Li, R., Xie, J., Zhou, X., Li, X., Liu, Q., Chen, X., Shen, W., Wang, L.: Federated learning-empowered smart manufacturing and product lifecycle management: A review. Advanced Engineering Informatics **65**, 103179 (2025) https://doi.org/10.1016/j.aei.2025.103179

[2] Purkayastha, A.A., Aggarwal, S.: Federated learning for predictive maintenance: A survey of methods, applications, and challenges. In: 2024 IEEE 67th International Midwest Symposium on Circuits and Systems (MWSCAS), pp. 238–242. IEEE, Springfield, MA, USA (2024). https://doi.org/10.1109/mwscas60917.2024.10658900 . http://dx.doi.org/10.1109/MWSCAS60917.2024.10658900

[3] Jia, N., Qu, Z., Ye, B., Wang, Y., Hu, S., Guo, S.: A comprehensive survey on communication-efficient federated learning in mobile edge environments. IEEE Communications Surveys and Tutorials **27**(6), 3710–3741 (2025) https://doi.org/10.1109/comst.2025.3535957

[4] Qian, C., Guo, Y., Lu, C., Yu, W.: Edge intelligence in smart manufacturing CPS, pp. 221–246. Elsevier, ??? (2025). https://doi.org/10.1016/b978-0-44-326572-3.00017-6 . http://dx.doi.org/10.1016/B978-0-44-326572-3.00017-6

[5] Artiushenko, V., Lang, S., Lerez, C., Reggelin, T., Hackert-Oschätzchen, M.: Resource-efficient edge ai solution for predictive maintenance. Procedia Computer Science **232**, 348–357 (2024) https://doi.org/10.1016/j.procs.2024.01.034

[6] Wong, K.-S., Nguyen, D.-M., Le, K., Ho, L., Do, C., Le-Phuoc, D.: An empirical study of federated learning on iot–edge devices: Resource allocation and heterogeneity. IEEE Transactions on Neural Networks and Learning Systems **37**(2), 753–765 (2026) https://doi.org/10.1109/tnnls.2025.3611415

[7] Ye, M., Fang, X., Du, B., Yuen, P.C., Tao, D.: Heterogeneous federated learning: State-of-the-art and research challenges. ACM Computing Surveys **56**(3), 1–44 (2023) https://doi.org/10.1145/3625558

[8] Huang, W., Ye, M., Shi, Z., Wan, G., Li, H., Du, B., Yang, Q.: Federated learning for generalization, robustness, fairness: A survey and benchmark. IEEE Transactions on Pattern Analysis and Machine Intelligence **46**(12), 9387–9406 (2024) https://doi.org/10.1109/tpami.2024.3418862

[9] Kisten, M., Ezugwu, A.E.-S., Olusanya, M.O.: Explainable artificial intelligence model for predictive maintenance in smart agricultural facilities. IEEE Access **12**, 24348–24367 (2024) https://doi.org/10.1109/access.2024.3365586

[10] Huong, T.T., Bac, T.P., Ha, K.N., Hoang, N.V., Hoang, N.X., Hung, N.T., Tran, K.P.: Federated learning-based explainable anomaly detection for industrial control systems. IEEE Access **10**, 53854–53872 (2022) https://doi.org/10.1109/access.2022.3173288

[11] Trockman, A., Kolter, J.Z.: Patches Are All You Need? (2022). https://arxiv.org/abs/2201.09792

[12] Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2018)

[13] Yuan, L., Wang, Z., Sun, L., Yu, P.S., Brinton, C.G.: Decentralized federated learning: A survey and perspective. IEEE Internet of Things Journal **11**(21), 34617–34638 (2024) https://doi.org/10.1109/jiot.2024.3407584

[14] Karami, M., Karami, A.: Harmony in federated learning: a comprehensive review of techniques to tackle heterogeneity and non-iid data. Cluster Computing **28**(9) (2025) https://doi.org/10.1007/s10586-025-05250-y

[15] Llasag Rosero, R., Silva, C., Ribeiro, B., Santos, B.F.: Label synchronization for hybrid federated learning in manufacturing and predictive maintenance. Journal of Intelligent Manufacturing **35**(8), 4015–4034 (2024) https://doi.org/10.1007/s10845-023-02298-8

[16] Davari, N., Veloso, B., Ribeiro, R.P., Pereira, P.M., Gama, J.: Predictive maintenance based on anomaly detection using deep learning for air production unit in the railway industry. In: 2021 IEEE 8th International Conference on Data Science and Advanced Analytics (DSAA), pp. 1–10. IEEE, Porto, Portugal (2021). https://doi.org/10.1109/dsaa53316.2021.9564181 . http://dx.doi.org/10.1109/DSAA53316.2021.9564181

[17] Veloso, B., Ribeiro, R.P., Gama, J., Pereira, P.M.: The metropt dataset for predictive maintenance. Scientific Data **9**(1) (2022) https://doi.org/10.1038/s41597-022-01877-3

[18] Davari, N., Veloso, B., Ribeiro, R., Gama, J.: MetroPT-3 Dataset. UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C5VW3R (2021)

[19] Sharma, A., Sim, K.Y., Chandrasekaran, S.: Fusionnet: intelligent sequence fusion for predictive maintenance in edge-enabled industrial iot systems. Cluster Computing **29**(10) (2026) https://doi.org/10.1007/s10586-026-06422-0

[20] Sharma, A., Sim, K.Y., Chandrasekaran, S.: Fusionnet lite: A lightweight federated deep learning model for privacy-preserving predictive maintenance (2026) https://doi.org/10.21203/rs.3.rs-8806936/v1

[21] C, H., D, S.V., Rajagopalan, N.: Federated learning and explainable ai-driven intrusion detection with hyperband optimization. Journal of Computer Virology and Hacking Techniques **21**(1) (2025) https://doi.org/10.1007/s11416-025-00571-3

[22] AI4I 2020 Predictive Maintenance Dataset. UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C5HS5C (2020)

[23] McMahan, B., Moore, E., Ramage, D., Hampson, S., Arcas, B.A.y.: Communication-Efficient Learning of Deep Networks from Decentralized Data. In: Singh, A., Zhu, J. (eds.) Proceedings of the 20th International Conference on Artificial Intelligence and Statistics. Proceedings of Machine Learning Research, vol. 54, pp. 1273–1282. PMLR, Fort Lauderdale, FL, USA (2017). https://proceedings.mlr.press/v54/mcmahan17a.html

[24] Sundararajan, M., Taly, A., Yan, Q.: Axiomatic attribution for deep networks. In: Precup, D., Teh, Y.W. (eds.) Proceedings of the 34th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 70, pp. 3319–3328. PMLR, Sydney, Australia (2017). https://proceedings.mlr.press/v70/sundararajan17a.html

## Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- noniidheterogeneity.csv
- significance.csv
- canonicalseedmetrics.csv
- igagreementseeds.csv