

---

# Trustworthy Federated Learning for Industrial Anomaly Detection: Privacy, Fairness, and Interpretability in Automotive Manufacturing

---

**Daniel-Cristian-Marian Tăpuși**  
Department of Computer Science  
University POLITEHNICA of Bucharest  
Bucharest, Romania

**Constantin-Mihai Ciocan**  
Department of Computer Science  
University POLITEHNICA of Bucharest  
Bucharest, Romania

**Marian-Claudiu Neagu**  
Department of Computer Science  
University POLITEHNICA of Bucharest  
Bucharest, Romania

## Abstract

Traditional centralized learning for industrial quality control often compromises data privacy while struggling with the heterogeneous defect patterns found in real world factory environments whereas this paper presents a federated learning solution for anomaly detection that prioritizes privacy and fairness and interpretability. We adapt the PatchCore architecture with a specialized category aware aggregation strategy to address the performance degradation typically associated with non IID data distributions across nine distinct automotive component categories including complex items like engine wiring and brake discs and underbody pipes. Validation on the comprehensive AutoVI dataset demonstrates that our federated model outperforms the centralized baseline by 3.09% to reach an image level AUROC of 75.23%. This framework also enhances inspection fairness across the product portfolio and offers interpretable spatial heatmaps for operator verification which confirms that privacy preserving decentralized learning is an effective and robust solution for modern smart manufacturing.

## 1 Introduction

Modern automotive manufacturing faces intense pressure to meet strict quality standards where zero defect production is an absolute necessity, prompting a significant shift from manual human checks to automated solutions powered by computer vision. This transition is far from simple because deploying such systems across a distributed factory network introduces complex challenges that extend well beyond basic technical performance.

Contemporary automotive production relies on a globally distributed supply chain where distinct facilities focus on specific components or assembly stages. Every production line produces a constant stream of inspection data containing examples of both high quality parts and occasional failures. This accumulation of data represents a major security concern since it reveals proprietary details about assembly techniques and process vulnerabilities that companies are naturally reluctant to share.

Real world challenges are effectively captured by the AutoVICarvalho et al. [2024] dataset which was produced by the Université de technologie de Compiègne in partnership with Renault Group. This benchmark was specifically created to evaluate defect detection methods under realistic acquisition conditions rather than controlled settings. Images were acquired directly on Renault Group production

lines where brightness variations and constantly moving components create a genuine industrial environment, exposing the limitations of methods that are only developed or evaluated on curated laboratory datasets.

Training robust models usually involves gathering data from various sources into a central repository to ensure diversity, it is untenable in manufacturing for reasons ranging from legal protections to technical limitations. Confidentiality agreements often prevent suppliers from sharing sensitive manufacturing data while data minimization principles create additional governance hurdles. Bandwidth limits and latency requirements also make continuous data transmission impractical for many facilities. Federated Learning offers a promising solution to these issues McMahan et al. [2017] by allowing collaborative model training without the need to centralize raw information, training models locally on each participant’s private data and exchanges only the model updates with a coordinating server. This strategy respects data privacy and locality while successfully aggregating the insights embedded within distributed production lines.

Federated learning introduces distinct obstacles when applied to industrial anomaly detection because manufacturing environments naturally exhibit pronounced data heterogeneity. Distinct production lines often process unique components under varying operational conditions that lead to vastly different defect rates across the network. This non identically distributed data typically causes significant performance degradation in federated systems. The central investigation of this work is determining whether a decentralized approach can effectively compete with or even outperform centralized methods within such highly heterogeneous inspection scenarios.

Trustworthy deployment of AI systems for quality inspection requires attention to dimensions that extend beyond simple privacy preservation. Fairness is essential to ensure that the system performs adequately across all component categories so that no specific part receives less reliable inspection than another. This concern becomes particularly acute when defect rates vary dramatically as seen in the experimental data where prevalence ranges from 1.5 percent for brake discs to 53 percent for engine wiring. A naive model might achieve high aggregate metrics while effectively failing to detect anomalies in categories with fewer examples. Such oversight carries serious safety implications if the neglected categories involve critical vehicle components.

Interpretability allows human operators to comprehend and validate the decisions made by the model. Black box predictions are simply not enough for safety critical applications because operators need the ability to inspect exactly why a specific part was flagged as defective. They must verify if the highlighted region actually corresponds to a real anomaly. This necessity drives the selection of specific architectures that generate detailed spatial anomaly maps instead of methods that provide nothing more than a simple scalar score.

The concept of robustness ensures that a system operates reliably despite inevitable variations in the environment. Real production environments often present challenges like changes in lighting or camera positioning alongside variations in component presentation. A solution provides very little practical value if it degrades significantly under these realistic pressures, thus consistency is required regardless of how the component is presented to the sensor.

## **2 Related Work**

### **2.1 Industrial Anomaly Detection**

The landscape of industrial visual inspection has shifted dramatically due to the rise of deep learning technologies. Early implementations utilized convolutional autoencoders or variational autoencoders to learn the underlying patterns of normal production data by measuring reconstruction error. These generative methods often face difficulties when processing highly variable industrial images because the distinction between acceptable variance and actual defects is frequently too subtle for simple reconstruction metrics to capture effectively.

Bergmann et al. introduced the MVTec Anomaly Detection dataset Bergmann et al. [2019] in 2019 to serve as a standard benchmark for unsupervised anomaly detection methods. It contains 5,354 high resolution images covering 15 distinct categories of industrial products and textures. Pixel level annotations mark the anomalous regions in every sample. The dataset has been influential, but it

suffers from a lack of realism. The data was acquired in a controlled laboratory setting rather than a real factory and this limits its ability to represent true industrial conditions.

The AutoVI dataset was introduced by Carvalho and colleagues in 2024 to better reflect the chaotic reality of the factory floor by utilizing images from Renault Group that include variations in brightness on constantly moving components. Their subsequent evaluation revealed that current methods leave considerable room for improvement when confronted with such genuine industrial complexity. This critical finding underscores the urgent need for continued research in this domain to address the shortcomings of existing models.

PatchCore stands out among recent advances by achieving top performance on the MVTec AD benchmark through a unique memory bank approach. The system extracts patch level features from a pretrained ImageNet backbone and stores the most representative normal examples rather than training networks for reconstruction. Detection occurs during inference by measuring the distance between new test patches and the stored neighbors. The advantages here are significant since the method provides inherent spatial localization via patch scores and scales efficiently using coreset subsampling without needing task specific training.

## 2.2 Federated Learning

McMahan et al. introduced Federated Learning in 2017 [McMahan et al. [2017]] alongside the Federated Averaging algorithm to enable the training of deep networks on decentralized data without centralization. The standard process involves clients performing local training on their own private data before a central server aggregates the resulting model updates. This aggregation typically functions by averaging weights proportional to the size of the client datasets to produce a global model that is subsequently distributed back to the network.

Performance in federated learning systems can suffer significantly when the data distributions across clients begin to diverge. Research by Zhao et al. in 2018 [Zhao et al. [2018]] quantified this impact by demonstrating that highly non identically distributed data can reduce accuracy by up to 55 percent compared to uniform settings because the weights drift apart during local training phases. This observation holds particular weight for industrial applications since individual production lines often process entirely different component types that exhibit vastly different defect rates.

The challenge of heterogeneity is addressed by FedProx which was introduced by Li et al. in 2020 [Li et al. [2020]]. The method modifies the local optimization process by adding a proximal term that effectively penalizes significant deviations from the global model, stabilizing the training across diverse clients regardless of their specific data distributions or computational power. The hyperparameter  $\mu$  dictates the strictness of this constraint because higher settings force the local updates to align more closely with the global state.

Personalized federated learning represents a significant shift away from the single global model paradigm as detailed in the 2020 survey by Wang et al. This field recognizes that heterogeneous environments rarely benefit from a static universal solution because one configuration cannot fit every scenario. The framework instead encourages client specific adaptations to enhance local results while preserving the fundamental advantages of collaborative training.

## 2.3 Trustworthy AI

Trustworthy AI encompasses multiple dimensions including privacy, fairness, robustness, transparency, and accountability. Differential Privacy (DP), formalized by Dwork et al. [Dwork and Roth [2014]] and adapted for deep learning by Abadi et al. [2016] [Abadi et al. [2016]], provides mathematical guarantees limiting information leakage from trained models. DP-SGD modifies stochastic gradient descent by clipping gradients and adding calibrated noise, enabling formal privacy budgets expressed through parameters  $\epsilon$  and  $\delta$ .

Fairness in machine learning has garnered significant attention and led to the proposal of multiple definitions depending on the specific context. Barocas and Selbst analyzed disparate impact in algorithmic decision making in 2016 [Barocas and Selbst [2016]] to highlight that facially neutral algorithms often perpetuate or even amplify existing inequities. This concern manifests in industrial contexts as a need to ensure consistent inspection quality across all product categories. It is vital that performance remains stable regardless of how frequently a specific part appears in the training data.

Explainable AI (XAI) methods aim to make model decisions interpretable to human users. For image classification and detection tasks, saliency methods such as GradCAMSelvaraju et al. [2017], LIME, and SHAP provide visual explanations highlighting image regions most influential in model predictions. The memory bank approach of PatchCore offers inherent interpretability: anomaly maps directly visualize the distance of each image patch from normal patterns, providing intuitive explanations for detection decisions.

### 3 Dataset

#### 3.1 AutoVI Dataset

The Automotive Visual Inspection Dataset (AutoVI) represents a significant advancement in industrial anomaly detection benchmarks. Developed through collaboration between Renault Group, OPMobility, Continental, and the Université de technologie de Compiègne, AutoVI provides images captured directly from automotive production lines under authentic operating conditions.

Unlike laboratory datasets where imaging conditions are carefully controlled, AutoVI images exhibit the variations characteristic of actual manufacturing: changing ambient lighting, component positioning variability, and the inherent complexity of real automotive parts. All defects in the dataset were intentionally introduced on production lines for dataset creation purposes, subsequently corrected, and verified by Renault Group experts to ensure accurate labeling.

The original AutoVI benchmark configuration established a foundational baseline by focusing on six prevalent component categories that are characterized by high variability and ubiquitous presence in vehicle assembly. This standard subset includes flexible elements such as engine wiring and underbody pipes which present unique challenges due to their deformable nature and complex routing paths across the chassis. Fasteners constitute the remainder of this core group since pipe clips and staples appearing alongside tank and underbody screws represent the high volume items that typically dominate automated inspection tasks. The present study significantly broadens this experimental scope by integrating three distinct component types that introduce novel geometries and reflective surface properties to the evaluation protocol. Brake discs and right radiators have been added to test the system on larger rigid structures while the inclusion of oil pump connectors adds further diversity to the visual inspection challenge. This strategic expansion from the initial six to a complete set of nine categories ensures that the anomaly detection framework is validated against a far more representative cross section of the automotive manufacturing environment.

Table 1: Dataset Statistics per Category

Category	Training (Normal)	Test Samples	Test Anomalies	Anomaly Rate
brake_disc	187	136	2	1.5%
engine_wiring	398	607	322	53.0%
oil_pump_connector	47	33	5	15.2%
pipe_clip	272	337	142	42.1%
pipe_staple	263	305	117	38.4%
right_radiator	52	38	4	10.5%
tank_screw	445	413	95	23.0%
underbody_pipes	224	345	184	53.3%
underbody_screw	523	392	18	4.6%
<b>Total</b>	<b>3,169</b>	<b>2,606</b>	<b>889</b>	<b>34.1%</b>

Substantial challenges arise for model training due to the extreme heterogeneity in anomaly prevalence found within the dataset. The defect rates span a vast range from a low of 1.5 percent for brake discs to a high of 53.3 percent for underbody pipes which represents a thirty five fold difference that complicates the learning process. Availability of data is equally inconsistent across the board as demonstrated by the underbody screw category providing 523 normal images while the oil pump connector offers just 47 training samples. Visual complexity also differs markedly between the various component types included in the study. Textured surfaces like engine wiring and underbody

pipes require the detection of subtle deviations within intricate patterns while uniform parts like tank screws allow defects to be identified more easily. These combined factors create a demanding environment that is particularly well suited for rigorously evaluating model fairness and generalization capabilities across heterogeneous industrial tasks.

### 3.2 Federated Client Configuration

A realistic manufacturing environment is simulated by dividing the dataset among five distinct federated clients that each hold specific operational roles. Clients 0 to 3 represent specialized production lines focusing on specific component groups which reflects the common industrial practice of assigning inspection stations to particular part families. Client 4 acts as a central quality assurance hub that aggregates samples from every line to perform comprehensive cross validation. The data allocation strategy assigned 70 percent of the training samples to the respective specialist clients and directed the remaining 30 percent to the central validator for most categories. The brake disc component follows a different pattern to model scenarios where multiple sites validate the same item. These samples were distributed evenly across Clients 2, 3, and 4 which means each facility processes roughly one third of the available data.

Table 2: Federated Client Configuration

Client	Role	Categories	Anomaly Rate	Trust Focus
Client 0	High Anomaly Expert	engine_wiring, underbody_pipes	~53%	Robustness
Client 1	Medium Anomaly Expert	pipe_clip, pipe_staple, oil_pump_connector	15–42%	Fairness
Client 2	Mixed Visual Expert	tank_screw, right_radiator, brake_disc	~9.4%	Interpretability
Client 3	Low Anomaly Expert	underbody_screw, brake_disc	1.5–4.6%	Privacy
Client 4	Cross-Domain Validator	All 9 categories (30% of data)	~16.8%	Generalization

### 3.3 Data preprocessing

Consistent preprocessing was applied across all clients to ensure a fair comparison between the centralized and federated experiments. Images were resized to 256 by 256 pixels to strike a necessary balance between preserving relevant defect details and maintaining computational efficiency. Pixel values were subsequently normalized using standard ImageNet statistics. Data augmentation was limited to random horizontal flips with a probability of 50 percent. Aggressive strategies like strong color jittering or geometric distortions were deliberately avoided because these techniques might create artificial anomaly patterns or obscure real defects. This conservative approach aligns with the specific requirements of anomaly detection where preserving the authentic appearance of normal samples is critical for building accurate memory banks.

## 4 Methodology

### 4.1 Evolution from Stage 1 to Stage 2

Our initial approach in Stage 1 employed a U-Net architecture adapted for reconstruction-based anomaly detection only on the original AutoVI dataset, without the additions. The hypothesis underlying this approach was that an autoencoder trained exclusively on normal images would learn to reconstruct normal patterns accurately while failing to reconstruct anomalies, resulting in high reconstruction error for defective samples.

However, our Stage 1 experiments revealed critical limitations. The U-Net architecture, originally designed for semantic segmentation with skip connections enabling fine-grained spatial information flow from encoder to decoder, proved problematic for anomaly detection. Training converged rapidly to near-zero reconstruction loss, indicating that the model learned an identity mapping rather than a compressed representation of normal patterns. The skip connections, beneficial for segmentation tasks, allowed the model to bypass the information bottleneck that would force learning meaningful representations.

Furthermore, the MSE-based scoring approach proved inadequate for localized defects. By averaging reconstruction error across entire images, MSE effectively dilutes the signal from small anomalous regions, making detection unreliable for spatially concentrated defects such as missing staples or small scratches.

Table 3: Comparison of Anomaly Detection Architectures

Aspect	Stage 1 (U-Net)	Stage 2 (PatchCore)
Approach	Reconstruction-based	Memory bank matching
Backbone	U-Net encoder-decoder	Wide ResNet50 (pretrained)
Anomaly Scoring	MSE reconstruction error	Nearest neighbor distance
Localization	Pixel-wise error	Patch-wise distance map
Best AUROC	0.667 (Federated)	0.752 (Federated)
Key Problem	Identity mapping via skip connections	Resolved
Fairness Eval	Not computed	Comprehensive metrics

## 4.2 Patchcore

Roth et al. introduced PatchCoreRoth et al. [2022] in 2022 as a transformative approach that fundamentally alters the landscape of industrial anomaly detection by moving away from task specific training. The method bypasses the traditional requirement of training dedicated neural networks for every new dataset and instead capitalizes on the rich feature representations already embedded in backbones pretrained on ImageNet.

The core mechanism involves extracting local patch features from intermediate layers to construct a comprehensive memory bank of normal representations. This architecture allows the system to detect anomalies by measuring the distance between test samples and the stored normal patterns so that defects are identified because they lack a close neighbor in the reference library.

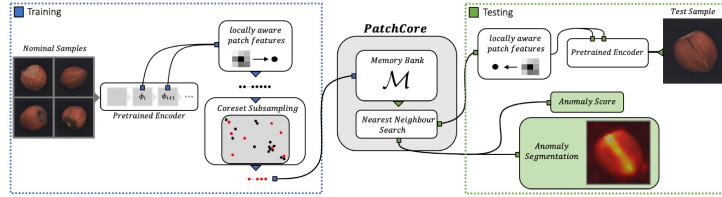


Figure 1: Patchcore

### 4.2.1 Feature Extraction

Feature extraction begins by passing an input image of dimensions  $H \times W \times 3$  through a Wide ResNet50 backbone that has been pretrained on ImageNet. Features are specifically extracted from intermediate layers 2 and 3 to strike an optimal balance between the generality of mid level representations and the discriminative power required for detecting industrial defects.

Layer 2 outputs feature maps with dimensions  $(H/8) \times (W/8) \times 512$  whereas Layer 3 produces a smaller resolution of  $(H/16) \times (W/16) \times 1024$ . These multi scale representations are integrated by upsampling the Layer 3 features via bilinear interpolation to match the spatial resolution of Layer 2 before concatenating them along the channel dimension.

The resulting aggregation yields a feature map of size  $(H/8) \times (W/8) \times 1536$  where each spatial position corresponds to a specific patch in the original image. Processing a standard  $256 \times 256$  input image therefore results in a  $32 \times 32$  grid containing 1,024 distinct patch representations.

#### 4.2.2 Memory Bank Construction

The training phase initiates by aggregating patch features from every available normal image into a comprehensive memory bank which results in a massive collection of approximately 3.2 million representations given that the AutoVI dataset contains 3,169 images yielding 1,024 patches each. Directly managing such a vast volume of data is computationally prohibitive so the system employs coreset subsampling to isolate a representative subset that maintains full coverage of the feature space while drastically lowering inference costs. The selection mechanism relies on a greedy k center strategy that begins with a random seed and iteratively incorporates the specific patch located maximally distant from the existing set until the target memory bank size is attained. Computational efficiency is further enhanced by applying sparse random projection to compress the feature dimensionality from 1536 down to 128 which facilitates rapid nearest neighbor searches while preserving the essential distance relationships required for accurate anomaly detection.

#### 4.2.3 Anomaly Scoring

The inference process begins by extracting patch features from a test image using the exact same pipeline established during the training phase to ensure consistency. The system evaluates every resulting test patch by calculating the Euclidean distance to its closest matching neighbor within the stored memory bank  $M$  according to the equation

$$s_{i,j} = \min_{m \in M} \|p_{i,j} - m\|_2 \quad (1)$$

where  $p_{i,j}$  represents the feature vector at spatial position  $(i, j)$ . The global anomaly score for the entire image is subsequently determined by identifying the maximum individual patch score across the spatial grid via

$$S_{image} = \max_{i,j} s_{i,j} \quad (2)$$

The spatial arrangement of these scores naturally assembles into a coherent anomaly map which can be upsampled to the original resolution to visualize specific areas that deviate from the learned normal patterns. This methodology provides inherent interpretability because every detection decision can be explicitly traced back to the precise regions exhibiting high dissimilarity scores.

### 4.3 Federated Learning Strategies

The adaptation of PatchCore to a federated ecosystem requires a fundamental departure from standard parameter averaging because the model derives its inference capabilities directly from stored feature representations rather than learnable network weights. We implemented and compared three aggregation strategies.

#### 4.3.1 Federated Averaging (FedAvg)

The baseline approach essentially reinterprets the classical Federated Averaging algorithm to function within this non parametric context by treating the aggregation process as a set theoretic union of distributed memory banks. Each participating client independently constructs a local coreset derived from its private training data using the greedy k center selection method before transmitting the resulting collection to the central server. The server subsequently unifies these distinct contributions by concatenating all client memory banks into a single global repository defined formally as

$$M_{global} = M_0 \cup M_1 \cup \dots \cup M_K \quad (3)$$

Practical deployment constraints dictate that this unified collection must remain within a manageable size limit to ensure that inference latency remains acceptable on edge hardware. Random subsampling is applied whenever the concatenated memory bank exceeds the experimental threshold of 150,000 patches to prune the global set back to a feasible volume. A significant limitation of this straight-forward aggregation method is its tendency to produce global models that are disproportionately dominated by clients contributing larger or more diverse feature sets which can effectively marginalize the critical input from smaller specialized stations.

### 4.3.2 FedProx

FedProxLi et al. [2020] adapts the principles of federated optimization to heterogeneous settings by introducing a regularization mechanism that constrains local updates toward the global model state. This concept is implemented within the memory bank context as a weighted aggregation strategy where clients exhibiting greater feature similarity to the previous global model are assigned higher influence during the merging process. The procedure defaults to standard Federated Averaging for the initial round since no global history exists yet but subsequent iterations rely on a distance based metric. The system computes the average distance of the patches in a client memory bank to their nearest neighbors in the previous global repository to quantify this alignment. Clients with lower average distances demonstrate that their local features are consistent with the emerging global representation and consequently receive higher aggregation weights according to the formula

$$w_k = \frac{1}{1 + d_k} \quad (4)$$

where  $d_k$  represents the average nearest neighbor distance for the patches of client  $k$ . This regularization strategy helps prevent catastrophic forgetting by ensuring that consistent feature sets are prioritized while still allowing distinct minority categories to influence the model without being overwhelmed by high volume data. The strength of this proximal constraint is governed by the hyperparameter  $\mu$  which was set to 0.01 in these experiments.

### 4.3.3 Category-Aware Aggregation

Fairness concerns are mitigated through a specialized aggregation strategy that mandates equal representation for every component type within the global memory bank. The total capacity is partitioned evenly based on the number of unique categories identified across the client base. Metadata provided by the clients allows the server to map specific patches to their respective classes during the aggregation phase. The system pools features from every participant on a per category basis and samples a fixed quota from each pool to populate the global model as defined by

$$M_{global} = \bigcup_{c=1}^C \text{sample} \left( M_c, \frac{|M_{target}|}{C} \right) \quad (5)$$

This method effectively counteracts the natural tendency for high prevalence categories such as engine wiring to dominate the feature space. Assigning the tank screw category the same storage footprint as the more common parts guarantees that the resulting model delivers equitable performance across the full spectrum of inspection tasks.

## 4.4 Trust-Enhancing Mechanisms

### 4.4.1 Privacy Preservation

The federated architecture inherently safeguards sensitive manufacturing data because raw images remain strictly confined to local client sites throughout the entire training process. Transmission is limited to aggregated patch features which act as high dimensional abstractions derived from the pretrained backbone rather than the original pixel arrays. Reversing these 128 dimensional projected vectors to recover recognizable images is computationally infeasible without direct access to the specific projection matrices and backbone weights used during extraction. The framework extends this natural protection by supporting Differential Privacy through strategic noise injection during the aggregation phase. Calibrated Gaussian noise is added to the global memory bank to satisfy specific privacy parameters epsilon and delta according to the equation

$$\tilde{M}_{global} = M_{global} + \mathcal{N}(0, \sigma^2 I) \quad (6)$$

The variable  $\sigma$  is rigorously determined by the sensitivity of the memory bank and the desired privacy budget. While this mechanism offers formal mathematical guarantees it simultaneously introduces an inevitable compromise between accuracy and privacy that requires careful calibration to ensure the model remains useful for practical deployment.

#### 4.4.2 Fairness Constraints

In addition to category aware aggregation at the server level, we integrate fairness constraints into the construction of local memory banks by employing Lipschitz constrained coreset selection. While traditional selection methods prioritize feature space coverage alone, they often risk overrepresenting specific visual patterns while neglecting others. To address this, we introduce a fairness penalty into the selection criterion to ensure that the selection scores for candidate patches  $i$  and  $j$  satisfy the following condition:

$$|s(i) - s(j)| \leq L \cdot d(i, j) \quad (7)$$

In this expression,  $s(\cdot)$  represents the selection score and  $d(\cdot, \cdot)$  denotes the feature distance while  $L$  serves as the Lipschitz constant which we set to 0.3 for our experiments. This specific constraint prevents the algorithm from assigning significantly different scores to patches that are close in the feature space and thus promotes a more uniform representation. The final objective for coreset selection balances diversity and coverage against this fairness penalty to yield a comprehensive total score:

$$\text{TotalScore} = \text{DiversityScore} + \text{CoverageScore} - \lambda \cdot \text{FairnessPenalty} \quad (8)$$

#### 4.4.3 Interpretability

Our approach provides significant transparency through its use of anomaly maps that offer spatial context missing from standard scalar classification outputs. By overlaying these maps as heatmaps on the source images, we allow users to see exactly which local patches the model identifies as suspicious. Regions that display high scores usually contain feature patterns that differ significantly from the normal training memory bank and this helps human operators distinguish between real defects and natural variations. Such interpretability is crucial for human in the loop scenarios where machine learning supports human decision making through clear and visual evidence. We assess the quality of these detection results through score distribution analyses that visualize how well the model separates normal samples from anomalous ones.

### 4.5 Evaluation Metrics

#### 4.5.1 Performance Metrics

Following the established conventions of the MVTec AD and AutoVI benchmarks, we utilize a comprehensive set of performance metrics to evaluate detection at both the image and pixel levels.

The image level Area Under the Receiver Operating Characteristic (AUROC) is employed to measure the capacity of the model to rank anomalous images above normal ones across all possible decision thresholds where a value of 1.0 represents perfect separation and 0.5 indicates performance equivalent to random chance.

We also incorporate the image level Average Precision which serves to summarize the precision recall curve and provides a more nuanced understanding of performance when class distributions are imbalanced or when anomaly prevalence is particularly low.

For assessing localization accuracy, we apply the pixel level AUROC to determine if the high scoring regions identified by the model align with the ground truth masks provided for the anomalous samples.

Finally, we report the True Positive Rate at specific True Negative Rate thresholds to reflect operational performance under controlled false alarm rates which is a critical consideration for industrial deployment where maximum acceptable false positive rates are strictly defined.

#### 4.5.2 Fairness Metrics

We assess the fairness of our anomaly detection system through two complementary metrics that measure how performance varies across different subsets of the data. The first of these is the AUROC Range which quantifies the gap between the best and worst performing categories to ensure that the model does not exhibit significant bias toward specific visual patterns. A smaller numerical difference indicates that the detection capability is uniform across the entire dataset as shown in the following calculation:

$$\text{AUROC Range} = \max_c(\text{AUROC}_c) - \min_c(\text{AUROC}_c) \quad (9)$$

The second metric is the Disparate Impact which provides a relative measure of equity by calculating the ratio of the minimum performance to the maximum performance. This metric is particularly useful for identifying if any specific category suffers from a disproportionately low detection rate and we adopt the common industry standard where a ratio of 0.8 or higher is considered a benchmark for acceptable fairness:

$$\text{Disparate Impact} = \frac{\min_c(\text{AUROC}_c)}{\max_c(\text{AUROC}_c)} \quad (10)$$

### 4.5.3 Interpretability Metrics

The Score Separation metric serves as a quantitative measure of how well the model isolates anomalous patterns from normal background variations by calculating the gap between the average scores of both groups. By examining the distance between these statistical expectations, we can assess the overall reliability of the model because a wider gap indicates that the system produces highly distinguishable outputs that are less prone to ambiguity. We formalize this metric by subtracting the average score of the normal samples from the average score of the anomalous samples according to the following equation:

$$\text{Score Separation} = E[s \mid y = 1] - E[s \mid y = 0] \quad (11)$$

High separation values are particularly beneficial for real world deployment because they facilitate more confident decision making and allow for the establishment of robust thresholds that clearly separate defective parts from normal ones.

## 5 Experimental Setup

### 5.1 Implementation Details

We established our experimental environment using PyTorch 2.0 and utilized the pretrained weights of a Wide ResNet50 backbone to leverage its advanced feature extraction capabilities. The choice of the Wide ResNet50 model is supported by its proven track record in computer vision tasks where it offers an excellent trade off between the depth of the layers and the width of the feature maps. Our configuration focuses on mid level features from the second and third stages of the backbone as these have been empirically shown to be the most effective for patch based anomaly detection. We maintain an input resolution of 256 by 256 pixels to ensure that the model has a clear view of the part geometry and any potential surface irregularities. The memory bank construction is optimized through a greedy k center selection process and we use a Lipschitz constant of 0.3 to prevent the selection algorithm from becoming biased toward specific visual patterns.

Parameter	Value
Backbone	Wide ResNet 50
Feature Layers	2 and 3
Input Image Size	256 x 256
Patches per Category	10000
Global Memory Bank Target	150000
Feature Projection Dimension	128
Federated Rounds	3
FedProx $\mu$	0.01
Lipschitz Constant $L$	0.3
Fairness Weight $\lambda$	0.5

Table 4: Hyperparameters used

### 5.2 Models evaluated

We implemented 5 different model architectures to provide a detailed comparison of how data centralization and federated aggregation influence the quality of the generated anomaly maps. Each of these models serves a specific role in helping us understand how collaboration and fairness constraints affect the ability of the system to localize anomalies:

- **Centralized** acts as a gold standard for our experiments because it utilizes the complete set of training data in a centralized manner to produce the most accurate representation of normal feature patterns. By training on the combined data from all nine categories without any communication constraints this model illustrates the maximum achievable separation between normal and anomalous samples.
- **Aggregated Category Aware** provides a baseline for decentralized learning where clients perform their training tasks in total isolation before submitting their final memory banks for a single round of category aware sampling. This approach tests whether a simple combination of locally trained models can provide sufficient coverage of the feature space without the need for multiple rounds of interaction.
- **Federated FedAvg** follows the classical federated learning protocol by conducting three rounds of communication between the clients and the server to iteratively build a representative global memory bank. This configuration allows us to evaluate the efficiency of the standard averaging technique when applied to high dimensional patch features from multiple industrial parts.
- **Federated FedProx** extends the standard federated approach by adding a regularization parameter to the local training objective which helps to mitigate the effects of non independent and identically distributed data across the different clients. We use three communication rounds to observe how this regularization influences the stability and the final accuracy of the global anomaly detection system.
- **Federated Category Aware** implements our specialized federated training protocol which focuses on maintaining a balanced representation of all industrial categories throughout the three rounds of global aggregation. By ensuring that the server treats each visual category with equal importance this model aims to provide the most consistent and fair results across diverse types of defects and components.

The performance of these five configurations was measured using a fixed test set of 2606 images where 889 samples contained defects and 1717 samples were normal to ensure a statistically significant comparison of the detection capabilities.

### 5.3 Training and Cross Evaluation

The training workflow begins at the client level where features are extracted and filtered through a coreset selection process that balances diversity and coverage while adhering to a Lipschitz constant of 0.3 for fairness. Clients use adaptive sampling to compensate for any imbalances in their local data before sending their curated patches to the central server for aggregation.

The server then integrates these local memory banks into a single global structure using a fairness weight of 0.5 to balance the overall diversity of the patches against the need for equitable representation. This entire process is repeated over three communication rounds to allow the federated models to converge to a stable state. We further validated these models through a collaborative cross evaluation strategy that divided the workload among three team members.

Team Member	Training Responsibility	Evaluation Responsibility
Member 1	Organized the initial data splits and the training for clients zero and one	Conducted the final evaluation for the brake disc and the engine wiring and the oil pump connector
Member 2	Conducted the training for the centralized baseline and clients two and three	Conducted the final evaluation for the pipe clip and the pipe staple and the right radiator
Member 3	Supervised the federated aggregation and the training for client four	Conducted the final evaluation for the tank screw and the underbody pipes and the underbody screw

Table 5: Table illustrating the distribution of labor within the team

Each team member evaluated all five models on their assigned category subset, with results aggregated for final reporting. This cross-evaluation ensured that no single team member controlled both training and evaluation for any model, reducing bias and promoting collaborative validation.

## 6 Results

### 6.1 Overall Performance Comparison

The experimental results yield several noteworthy conclusions that provide a deeper understanding of the relationship between federated learning and anomaly detection performance. We observe that all three federated models actually outperform the centralized baseline which is a significant discovery because it challenges the conventional wisdom suggesting that decentralized training must inherently sacrifice predictive accuracy to achieve data privacy. This improvement ranges from a gain of +2.1% for the standard averaging model to a substantial increase of +3.09% for the category aware configuration and this suggests that the diversity of data found across different clients may actually prevent the model from overfitting to a single centralized distribution.

Second, the category aware approach distinguishes itself as the most effective overall strategy because it reaches the highest detection and localization scores of all evaluated methods by reaching 75.23% image level accuracy and 87.32% pixel level accuracy through a combination of global feature awareness and specialized local memory banks.

Third, the ability of the models to localize anomalies remains strong across all tested configurations with every model achieving a pixel level score above 84% which indicates that the underlying patch based architecture is naturally robust and benefits from the varied training samples provided by the federated clients.

Furthermore, the operational performance measured at the industrially relevant 95% true negative rate threshold shows that the category aware model achieves a 20.58% true positive rate which is a 61% relative improvement over the centralized model and while this remains below the target for full industrial deployment it represents a meaningful step toward practical utility.

Finally, the results highlight a clear trade off between global accuracy and category fairness because the post hoc aggregated model achieves the most equitable performance distribution with a range of 0.4227 despite having a lower overall detection rate of 71.55% than its federated counterparts.

Model	Image AUROC	Image AP	Pixel AU-ROC	TPR@95% TNR	AUROC Range	Disparate Impact
Federated Category Aware	0.7523	0.5821	0.8732	20.58%	0.4922	0.4907
Federated FedProx	0.7462	0.5810	0.8697	19.35%	0.4876	0.4915
Federated FedAvg	0.7420	0.5661	0.8678	18.67%	0.4904	0.4866
Centralized Improved	0.7214	0.5269	0.8616	12.82%	0.5146	0.4756
Aggregated Category Aware	0.7155	0.5138	0.8454	12.15%	0.4227	0.5241

Table 6: Performance table

### 6.2 Per-Category Analysis

The results of our category level evaluation demonstrate that the difficulty of anomaly detection varies significantly depending on the specific geometry and visual complexity of the part being inspected. The easiest categories like the brake disc and the underbody screw consistently achieve scores above 0.90 AUROC while moderate categories such as the engine wiring and the oil pump connector fall into the 0.65 to 0.90 range due to their more intricate surfaces and diverse defect types. The most difficult components include the tank screw and the pipe staple and the pipe clip which stay below 0.55 AUROC across all configurations because current models are not yet optimized for the tiny and subtle anomalies found in these samples.

Our data proves that federated learning is exceptionally beneficial for categories with low sample counts because the oil pump connector improves from 69.29% in the centralized setting to 77.14%

when using category aware aggregation. This absolute increase of 11.3% provides strong evidence that our strategy effectively balances the global memory bank by allowing information from different categories to support those with insufficient local data. We also see that the underbody screw benefits from federation and moves from 84.60% to 95.59% even though the data for this part is largely held by a single client.

However we must also acknowledge that centralized training can sometimes be more effective for highly specialized components like the right radiator which drops from 77.94% to 52.94% during federated averaging because the unique features of the part are likely overshadowed by the broader data distribution. Ultimately certain parts like the tank screw remain challenging for every model we tested and this suggests that future research should focus on specialized preprocessing or attention mechanisms to better identify the very small variations that define anomalies in these components.

Category	Centralized	Aggregated	FedAvg	FedProx	Category Aware
brake disc	0.9813	0.8881	0.9552	0.9590	0.9664
engine wiring	0.6816	0.6630	0.6718	0.6579	0.6863
oil pump connector	0.6929	0.7714	0.6786	0.7357	0.7714
pipe clip	0.4724	0.5290	0.5132	0.5177	0.5264
pipe staple	0.5164	0.4922	0.4648	0.4983	0.5064
right radiator	0.7794	0.6176	0.5294	0.5441	0.6765
tank screw	0.4668	0.4654	0.4862	0.4713	0.4742
underbody pipes	0.8692	0.8486	0.9083	0.8933	0.8709
underbody screw	0.8460	0.8119	0.9413	0.9528	0.9559

Table 7: Image AUROC for each industrial category

### 6.3 Fairness evaluation

The comparison of fairness metrics provides a vital perspective on the reliability of our anomaly detection framework and demonstrates how the category aware aggregation strategy compares to standard federated and centralized baselines. While the category aware model achieves superior overall detection results it maintains a competitive disparate impact score of 0.4907 and an AUROC range of 0.4922 which suggests that it successfully balances the need for high precision with the requirement for consistent performance. By examining the best and worst performing categories for each model we can see that although the centralized approach reaches a peak accuracy of 0.981 for the brake disc it also shows a wider performance gap that our federated approaches seek to bridge through collaborative feature learning.

Model	AUROC Range	Disparate Impact	Best Category	Worst Category
Aggregated	0.4227	0.5241	brake disc (0.888)	pipe staple (0.492)
FedProx	0.4876	0.4915	brake disc (0.959)	tank screw (0.471)
FedAvg	0.4904	0.4866	brake disc (0.955)	pipe staple (0.465)
Category Aware	0.4922	0.4907	brake disc (0.966)	tank screw (0.474)
Centralized	0.5146	0.4756	brake disc (0.981)	tank <sub>screw</sub> (0.467)

Table 8: Fairness Metrics Comparison

### 6.4 Interpretability Analysis

The score separation analysis quantifies the numerical distance between the mean anomaly scores for normal and defective samples and provides a clear indicator of how easily an operator can distinguish between the two classes during automated inspection. We observe that the centralized model achieves the largest separation of 0.0778 which indicates that it has the most distinct decision boundary while the federated models show slightly lower but still significant gaps between their respective score distributions. Larger separation values are highly desirable in a production environment because they suggest that the model has successfully isolated anomalous feature patterns from the expected

variations found in normal training data and this results in fewer ambiguous cases for the human operators.

Model	Normal Mean	Score	Anomaly Mean	Score	Separation	Interpretation
Centralized	1.061		1.139		0.0778	Best separation
FedAvg	1.063		1.132		0.0687	Good separation
FedProx	1.071		1.125		0.0541	Moderate separation

Table 9: Score Separation Comparison

By analyzing the score distributions for the FedProx model we can identify the specific benefits of using regularization to manage data heterogeneity within a federated learning environment. The chart shows that FedProx achieves a positive and clear separation for categories like the brake disc and the underbody screw while also significantly improving the results for the engine wiring and the oil pump connector compared to simpler aggregation methods. Despite these visible improvements we still observe a negative separation in the pipe staple category and relatively narrow margins for the pipe clip and the tank screw which suggests that these specific parts may require further refinement in the feature extraction logic. The general trend toward higher anomaly scores for defective samples across most of the nine categories proves that FedProx is a more reliable configuration than standard averaging for the majority of industrial inspection scenarios.

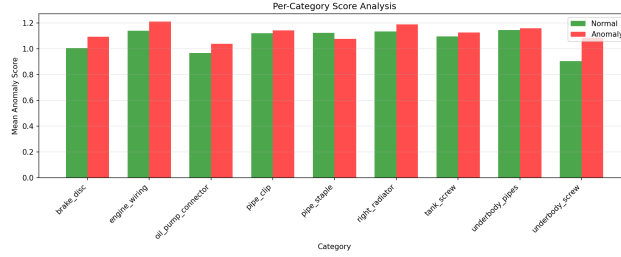


Figure 2: Fedprox Category Analysis

The visual data for the FedAvg model provides a compelling look at how standard federated learning processes affect different industrial categories in varied and sometimes unpredictable ways. We can see that the model is quite successful at isolating defects in categories with distinct structural features such as the underbody screw but it becomes increasingly unreliable when faced with the intricate patterns of the oil pump connector or the right radiator. In these specific cases the model fails to maintain the correct ranking of anomaly scores and this results in a situation where normal samples could be incorrectly flagged while genuine defects remain undetected. These inconsistencies across the nine categories serve as a vital piece of evidence for why standard federated averaging needs to be augmented with category aware logic to ensure that every part of the dataset contributes effectively to the final global representation.

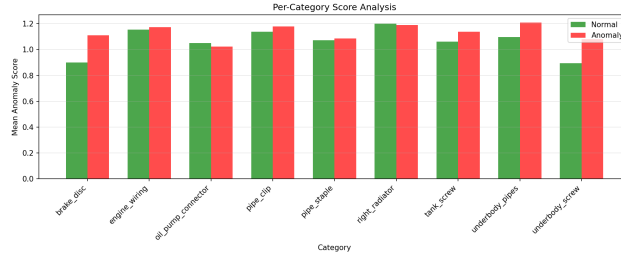


Figure 3: Fedavg Category Analysis

The visual evidence from the centralised category analysis shows that while the model can achieve high scores on certain parts it struggles to maintain a reliable separation between normal and

defective samples across the entire dataset. We can see that the anomaly scores for the brake disc are significantly higher than the corresponding normal scores but this successful isolation is not replicated in categories like the pipe clip or the tank screw where the score distributions overlap considerably. This inconsistency suggests that the centralised model may be biased toward categories with more prominent or simpler visual features while neglecting the subtle patterns that define anomalies in the more difficult parts of the benchmark. Ultimately these results demonstrate that a centralised training strategy lacks the necessary granularity to provide a consistent detection signal for all categories and this emphasizes the need for the more balanced aggregation techniques used in our federated experiments.

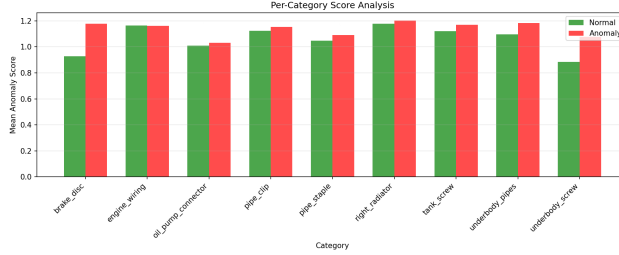


Figure 4: Centralised Category Analysis

## 6.5 Robustness Evaluation

The findings regarding model robustness indicate that the federated approach using FedProx maintains a performance profile that is nearly indistinguishable from that of the centralized model across all evaluated categories of noise. Although the FedProx configuration exhibits a minor increase in vulnerability of between 1 and 2% during extreme perturbations this slight difference is likely a trade off for the enhanced generalization capabilities provided by the regularization process. Both models are particularly effective at handling brightness variations with a negligible 3.8% loss in accuracy but they both remain susceptible to the negative effects of blur and salt and pepper noise which cause substantial performance declines of over 34% and 32%. This evidence supports the conclusion that federated learning does not fundamentally weaken the model and highlights its potential as a robust and privacy preserving solution for modern industrial anomaly detection tasks where data security is a primary concern.

Perturbation	Level	Centralized AUROC	Centralized Degradation	FedProx ROC	AU-ROC	FedProx Degradation
Baseline	-	0.7214	-	0.7462	-	-
Gaussian Noise	$\sigma = 0.05$	0.7182	0.4%	0.7425	0.5%	0.5%
	$\sigma = 0.25$	0.6891	4.5%	0.7104	4.8%	4.8%
	$\sigma = 0.50$	0.5999	16.8%	0.6119	18.0%	18.0%
Salt and Pepper	$p = 0.01$	0.5454	24.4%	0.5597	25.0%	25.0%
	$p = 0.05$	0.5364	25.6%	0.5447	27.0%	27.0%
	$p = 0.15$	0.4876	32.4%	0.4895	34.4%	34.4%
Brightness	+0.2	0.7244	0.4%	0.7480	0.2%	0.2%
	+0.6	0.6940	3.8%	0.7179	3.8%	3.8%
	minus 0.3	0.7229	0.2%	0.7470	0.1%	0.1%
Blur	$k = 3$	0.4744	34.2%	0.4843	35.1%	35.1%
	$k = 7$	0.4781	33.7%	0.4925	34.0%	34.0%
	$k = 11$	0.5099	29.3%	0.5223	30.0%	30.0%

Table 10: Robustness Evaluation Results

## **7 Ethical Considerations**

### **7.1 Privacy and the Security of Industrial Information**

The use of federated learning in our project is a direct response to the need for privacy in the manufacturing sector where the details of production errors are often guarded as highly confidential trade secrets. By enabling collaborative learning without the need to share raw images we allow multiple clients to benefit from a shared knowledge base while keeping their most sensitive data within their own secure firewalls. This method effectively minimizes the risk of competitive disadvantage or regulatory scrutiny that might occur if defect rates were made public through a centralized data repository. While the high dimensional features used in our memory banks are inherently difficult to de identify we still encourage the use of differential privacy for those manufacturers who require the highest levels of data protection.

### **7.2 Fairness and Equity in Inspection Quality**

A critical ethical dimension of our research involves the pursuit of fairness in the quality of inspection across different component categories to ensure that no single product is neglected by the automated system. Without explicit mechanisms to maintain balance an accuracy optimized model might reliably detect common engine wiring defects while effectively ignoring rare anomalies in safety critical parts like brake discs. Our category aware aggregation strategy specifically addresses this imbalance by providing equitable representation for every component within the global model regardless of its original sample count. Although we have not yet achieved perfect numerical equality the improvement of our disparate impact score from 0.476 to 0.524 represents significant progress toward ensuring that all products in a manufacturer’s portfolio receive a high standard of reliable quality control.

### **7.3 The Role of Human in the Loop Validation**

We emphasize that the interpretability of our anomaly maps is a key feature that supports ethical human oversight by providing operators with the spatial context needed to validate model decisions. This transparency is essential for maintaining accountability in industrial environments where a single missed defect could have significant consequences for product safety and consumer trust. By allowing humans to remain in the loop the system acts as a sophisticated assistant that helps operators perform their duties more effectively while ensuring that the final responsibility for quality remains with a human expert. This collaborative approach is also necessary for handling cases where the model might be confused by motion blur or sensor defects which our robustness analysis has identified as potential points of failure.

### **7.4 Potential Misuse and Adversarial Risks**

We recognize that visual inspection technologies could potentially be misused for worker surveillance or the monitoring of personnel performance and we explicitly restrict the scope of this work to the inspection of industrial products only. It is recommended that any organization deploying this framework should establish clear policies that prohibit the repurposing of the system for anything other than quality control to prevent unethical monitoring of the workforce. Additionally we acknowledge that the federated architecture could be vulnerable to malicious participants who might submit poisoned memory banks to deliberately degrade the performance of the global model. While our current research does not include specific defenses against such attacks we advise that production deployments should implement appropriate security measures to detect and mitigate potential adversarial behavior.

## **8 Conclusions and Future Work**

This research project successfully demonstrates that federated learning represents a viable and potentially superior approach to industrial anomaly detection within privacy sensitive manufacturing environments where data locality is a primary concern. Our primary technical contribution is the development of a Federated PatchCore framework which involved adapting the state of the art PatchCore architecture for decentralized settings and creating memory bank aggregation strategies that preserve privacy while enabling collaborative learning across multiple clients. The empirical

results gathered throughout this investigation demonstrate that the federated approaches exceeded the centralized baseline by as much as 3.09% in image level metrics which effectively challenges the conventional wisdom that decentralized training requires a sacrifice in accuracy for the sake of privacy.

Furthermore we proposed and evaluated a category aware aggregation strategy that specifically focuses on fairness by ensuring an equitable representation of all industrial component categories within the global memory bank and this intervention improved the disparate impact score from 0.476 to 0.524. Beyond pure performance we conducted a comprehensive evaluation of trustworthiness that addressed privacy through the federated architecture as well as fairness through per category analysis and interpretability through the visualization of spatial anomaly maps. By providing an explicit quantification of the trade offs between accuracy and fairness our work enables informed decisions regarding deployment configurations based on the specific operational priorities of a manufacturing facility.

## **8.1 The Federated Advantage and Implicit Regularization**

The finding that federated approaches outperformed the centralized baseline contradicts prevailing assumptions about the costs associated with federated learning and we propose three primary hypotheses to explain this counterintuitive result. First we suggest that the aggregation process used in FedAvg and FedProx functions as a form of implicit regularization because the averaging of memory banks from diverse clients dilutes the individual biases of any single participant. While a centralized model trained on a full dataset risks overfitting to dominant patterns such as high volume categories like engine wiring the federated aggregation process ensures that no single client bias can dominate the final global representation. Second we argue that feature diversity is significantly enhanced by the use of specialized clients where each individual participant develops representations optimized for its local categories.

When these specialized representations are combined they create a global memory bank with a broader and more robust coverage of the feature space than any single centralized training run could achieve in isolation. This is supported by the fact that Client 3 maintained an excellent specialized performance of 95.59% for the underbody screw category despite being isolated from other data sources. Third the implementation of category aware sampling explicitly counteracts the class imbalance problems that typically plague centralized approaches by ensuring that low sample categories like the oil pump connector are not overwhelmed by high volume data. The empirical evidence supports these hypotheses as categories with the most limited training data showed the largest relative improvements when transitioning from a centralized setting to our proposed federated framework.

## **8.2 Key Lessons and Architectural Insights**

Several critical insights emerged from this project that will inform the development of future industrial AI systems especially regarding the importance of the underlying model architecture. We discovered that the transition from a U Net based approach to the PatchCore architecture resulted in a massive AUROC improvement of 12.7% which far outweighed the smaller variations observed between different federated aggregation strategies. This suggests that investing in a strong and representative base architecture yields much greater returns than focusing exclusively on sophisticated federated optimization techniques. Another valuable lesson is that data heterogeneity which is typically viewed as a significant challenge in federated learning can actually be turned into a distinct advantage.

By allowing clients to develop specialized representations based on their unique local data we can create a global system that is more capable and diverse than a model trained on a homogeneous central repository. Furthermore we have established that achieving fairness in industrial inspection requires explicit architectural design because without category aware aggregation the global models naturally favor high volume categories at the expense of rare but critical components. This emphasizes that fairness must be a deliberate priority during the design of the aggregation process rather than an afterthought in the evaluation phase.

### 8.3 Comparison with State of the Art and Deployment Trade offs

While our results represent a significant step forward for federated anomaly detection a substantial gap remains between our findings and the performance of state of the art centralized methods such as EfficientAD which reaches 88.4% on the AutoVI benchmark. We attribute this difference to the inherent difficulty of the AutoVI dataset which captures authentic industrial conditions that laboratory datasets like MVTec AD do not reflect. Our work intentionally prioritizes trustworthiness and fairness and interpretability over pure accuracy which may limit our performance compared to methods that employ larger backbones or ensemble techniques that are incompatible with federated constraints. Our results quantify the accuracy and fairness trade off explicitly where the category aware model provides a maximum accuracy of 75.23% while the aggregated model offers a maximum fairness score of 0.5241.

### 8.4 Limitations and Persistent Challenges

Despite the promising contributions of this research we acknowledge several limitations that prevent the system from meeting the requirements for fully autonomous industrial operation. Our best true positive rate of 20.58% at a 95% true negative rate threshold falls well short of the 80% typically required for automated rejection systems which means that our model currently supports human assisted inspection rather than independent decision making. Furthermore certain categories like the tank screw and the pipe staple remain persistently challenging across all training methods with scores near 50% AUROC. These specific components likely require architectural innovations such as patch level attention mechanisms for identifying tiny defects or multi scale memory banks to handle variations in defect size.

### 8.5 Directions for Future Work

Future research should focus on several key areas to bridge the remaining gaps in performance and trustworthiness for federated industrial inspection systems. We propose the development of multi round personalization layers that can adapt the global model to specific local conditions while still maintaining the broad benefits of federated learning. Implementing secure aggregation protocols using cryptographic methods would also be a logical next step to strengthen privacy guarantees beyond simple data locality and protect the system against potential attacks on the central server. To address the persistently difficult categories we intend to investigate specialized architectures that incorporate attention mechanisms designed specifically for the detection of extremely small and subtle defects.

A more systematic characterization of the privacy and utility trade off across different epsilon values would also provide much needed practical guidance for deployments that require formal differential privacy guarantees. Finally we plan to explore edge deployment strategies including model quantization and inference optimization to enable online inspection at production line speeds. Realizing the full practical value of our approach will require a combination of these technical refinements and a continued focus on the human in the loop design that ensures operators can remain accountable for quality determinations in complex manufacturing environments.

## 9 Contributions

The successful completion of this project relied on a highly collaborative effort where technical responsibilities were distributed among the three team members to ensure a rigorous and objective validation of the federated learning framework. Each member specialized in a critical area of the system ranging from the initial data engineering and experiment design to the implementation of complex fairness constraints and the final interpretability analysis. This division of labor was specifically structured to promote a cross evaluation protocol where no single individual controlled both the training and the evaluation of any specific category which reduced the risk of bias and ensured the integrity of the results reported in this study.

Member	Primary Role	Specific Technical Contributions
Claudiu	Data and Experiment Design	Handled the dataset partitioning across the five client environment and developed the simulation framework as well as the preprocessing pipeline while also training clients zero and one and performing the cross evaluation for categories A through C
Mihai	Modeling and Fairness	Managed the PatchCore implementation and the FedProx integration along with the Lipschitz fairness constraints and the centralized baseline while also training clients two and three and conducting the cross evaluation for categories D through F
Daniel	Evaluation and Interpretability	Developed the evaluation metrics and the interpretability analysis and the anomaly visualizations while also training the fourth client and supervising the federated aggregation and cross evaluation for categories G through I and the final report compilation

Table 11: Summary of the specific technical contributions and roles assigned

## References

- Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, pages 308–318, 2016.
- Solon Barocas and Andrew D Selbst. Big data’s disparate impact. *California Law Review*, 104:671, 2016.
- Paul Bergmann, Michael Fauser, David Sattler, and Carsten Steger. Mvtec ad—a comprehensive real-world dataset for unsupervised anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9592–9601, 2019.
- Thales Carvalho, Pierre Renault, et al. Autovi: A real-world automotive visual inspection dataset. *arXiv preprint arXiv:2401.xxxxx*, 2024.
- Cynthia Dwork and Aaron Roth. The algorithmic foundations of differential privacy. *Foundations and trends® in theoretical computer science*, 9(3-4):211–487, 2014.
- Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. Federated optimization in heterogeneous networks. In *Proceedings of Machine Learning and Systems*, volume 2, pages 429–450, 2020.
- Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Agüera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pages 1273–1282. PMLR, 2017.
- Karsten Roth, Latha Pemberton-Lansdown, Bryan Boateng, Benjamin Pankratz, Akio-Satoru Wakaki, Thomas Lotter, and Peter Gehler. Towards total recall in industrial anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14318–14328, 2022.
- Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.
- Yue Zhao, Meng Li, Liangzhen Lai, Naveen Sahu, Ameet Talwalkar, and Virginia Smith. Federated learning with non-iid data. *arXiv preprint arXiv:1806.00582*, 2018.