# Uncertainty-Guided Bandwidth-Adaptive Model Compression for Industrial Pore Detection

Prosenjit Roy1, Kijoon Lee2, Mohsen Taheri Andani2, Noushin Ghaffari1 1Department of Computer Science, Prairie View A&M University, Prairie View, TX, USA
2Department of Mechanical Engineering, Texas A&M University, College Station, TX, USA
Email: 1prosenjit@student.pvamu.edu, nghaffari@pvamu.edu, 2kijoonlee@tamu.edu, mtaheri@tamu.edu

***Abstract*— Manufacturing systems require accurate pore detection in products. Edge computing provides benefits but creates bandwidth limitations. Current compression methods ignore network conditions. They also lack confidence scores for critical decisions.**

**This paper presents a new compression framework. Uncertainty guides decisions. The method adapts to bandwidth changes automatically. The system picks the best model size based on network speed and prediction confidence. Monte Carlo dropout provides uncertainty values. The framework balances accuracy with communication costs.**

**Evaluation uses 720 industrial pore samples. Tests cover four compression levels: 1.0, 0.75, 0.50, and 0.25. The method achieves 0.670 F1 score. This represents 17.8% improvement over fixed compression. Communication costs drop by 20%. The optimal model shrinks by 43.7% to 9.29MB. Accuracy remains at 99.54%. Uncertainty values range from 0.0014 to 0.0020. Values correlate with compression levels in predictable ways.**

**Quality control workers can use uncertainty maps for decisions. The framework solves real deployment problems. The system meets both efficiency and reliability needs.**

***Index Terms*— Bandwidth adaptation, Edge computing, Industrial vision, Neural network compression, Pore detection, Quality control, Uncertainty quantification**

## I. Introduction

Industrial pore detection has become critical for quality control in manufacturing processes. Pores in materials like aluminum alloy die-cast plates can lead to structural failures. In addition, accurate local detection is essential in the context of safety-critical applications [1]. Existing inspection approaches falter in the presence of noise, changes in lighting and textures. Deep learning methodologies solve these issues however suffer from deployment issues in low-power settings [2].

Edge computing deployment is advantageous for manufacturing vision systems in terms of low latency and reliability. However, edge devices are computationally restrictive, thus cannot handle complex models. Model complexity can be further reduced without loss of performance using model compression strategies such as pruning, quantization, and knowledge distillation [3]. Cloud offloading supplies computing capability and high-volume storage, but has a bandwidth constraint. Current compression techniques discriminate the network conditions during the link transmission. Pore detection presents special challenges. These challenges motivate the proposed approach. First, labeled defect data costs money. Small datasets are common. This makes overfitting a real problem. Second, missing a defect can cause failures. Buildings or planes might break. Confidence scores are necessary for safety. Third, factories have different internet speeds. Some production lines have fast networks. Others have slow ones. Models must adapt. Fourth, human inspectors need help. They want to know where to look carefully. These needs differ from normal computer vision. They require uncertainty measures combined with smart deployment. Recent advances in model compression focus primarily on reducing model size or computation. LayerCollapse reduces model depth through structured pruning [4]. ChannelZip offers task-adaptive compression for IoT devices [5]. However, these methods do not provide a measure of uncertainty about the deployment decision. Uncertainty quantification is necessary in safety-critical applications where the model trust impacts on the decision of operating [6]. Bandwidth-aware systems tune processing to the network conditions. The techniques are amended to adjust the compression ratio based on the state of the communication [7]. In image analytics offloading, compression should be taken into consideration, including weight flow [8]. However, most current bandwidth adaptation mechanisms mainly aim at optimizing the communication, not the model selection with uncertainty consideration.

Industrial pore detection presents unique challenges requiring both accuracy and reliability. Pore formation is a relatively rare event, and imbalance issues exist in the dataset [9]. On the other hand, the early defect detection is important to avoid the build failures adopted for the manufacturing [10]. Existing methods focus on either accuracy (or sectors of the data) without considering efficiency, and this paper have effectively developed a method for generating uncertainty-informed deployment decisions for critical situations that have not previously been investigated.

This paper presents a framework combining uncertainty with bandwidth adaptation. The system targets industrial pore detection specifically. The main contributions are: First, the framework combines Monte Carlo dropout with model selection. This combination is new for industrial defect work. This integration has not been done before for quality control. Second, the evaluation tests all compression levels carefully. Results show how uncertainty changes with compression. The patterns are clear. Third, the method achieves 17.8% better

F1 scores than fixed approaches. Communication costs reduce by 20%. Fourth, the framework provides a real deployment system. The system gives confidence scores that safety teams can use.

## II. Related Works

### A. Model Compression Techniques

Pore formation events occur at low frequency and imbalanced data problem is common [9]. On the other hand, early defect discovery is important to avoid build failures in production [10]. Operators of quality control require measures of confidence for decision support. Existing approaches consider accuracy or efficiency only separately, rather than considering both integrated uncertainty-guided depl0yment. Neural network compression addresses deployment constraints through various strategies. Structured pruning methods remove entire network components to reduce computational overhead. Layer collapse introduces regularization to encourage shallow fully connected layers thus allowing for training and compressing without the need for further fine-tuning [4]. Channel-wise pruning methods aim to eliminate uninformative feature maps. ChannelZip adapts shapley value in evaluating the importance of a channel and works SLAs in its compression decisions [5].

Quantization techniques lower numerical precision at the cost of preserving model performance. Recent developments mitigate the overhead of quantization-aware training for efficient deployment [3]. In knowledge distillation, knowledge is distilled from large teacher models into small student networks. PQK is a pruning-quantization-knowledge distillation based approach for edge devices aware compression [11]. This knowledge distillation based on Tucker decomposition could effectively transfer knowledge with less computing complexity [12]. However, these methods make only static compression decisions and do not consider deployment settings and do not provide confidence estimates for model predictions.

### B. Uncertainty Quantification in Edge Computing

Uncertainty estimation is particularly relevant for safety-critical contexts. Monte Carlo dropout provides an empirical uncertainty quantification by treating dropout as a bayesian approximation [13]. Recent studies show that both optimization and loss functions that exploit the predictions quantifications make Monte Carlo dropout more efficient [14]. Memristor-based Bayesian networks provide extreme-edge uncertainty quantification directly in the hardware [6]. MC-CP leverages Monte Carlo dropout and conformal prediction for scalable robust uncertainty quantification [15]. Applications in the context of time series show the efficacy of Monte Carlo dropout for uncertainty estimation in ESNs [16]. Related research includes uncertainty quantification in manufactur ing scenarios. However, previous uncertainty estimation strategies are designed to increase the accuracy of predictions, not to drive deployment decisions on limited resources.

### C. Bandwidth-Adaptive Systems

Communication-aware systems are those that are optimized for performance given network constraints. Federated learning dynamically adjusts the compression ratio according to bandwidth and communication state [7]. The ACFL algorithm adaptive compresses ratios on network conditions so as to ensure convergence. Gradient compression approaches trade off the communication efficiency and convergence of the model in distributed training [17].

Image compressor systems for bandwidth-limited offload. Context-aware optimization frameworks adapt their compression strategies according to the network state and the application demand [16]. Federated edge learning knowledge distillation reconciles heterogeneous device capabilities and keeps the models' performance [18]. But these works more on the optimization of communication than on model selection-based decision in the presence of uncertainty about deployment while node selects appropriate low power communication strategy.

## III. Comparison With Existing Approaches

The proposed approach differs from existing methods in key areas. Specifically, the framework combines uncertainty estimation with adaptive bandwidth control. This dual capability is not commonly found in prior work, making the system well-suited for industrial environments where real-time decision-making must be both efficient and reliable [18].

Table I outlines the major capabilities of various existing approaches compared to the proposed method.

TABLE I: Capabilities of Different Methods

| Method | Has Uncertainty | Adapts to Network | Safety Focus | Application Domain |
|---|---|---|---|---|
| LayerCollapse | No | No | No | General use |
| ChannelZip | No | Partly | No | IoT devices |
| PQK | No | No | No | Edge devices |
| ACFL | No | Yes | No | Training only |
| Proposed method | Yes | Yes | Yes | Factory quality |

LayerCollapse works by removing layers to compress models, while ChannelZip selects important channels during inference. Though these methods improve model size or efficiency, they do not consider prediction confidence, which is crucial in high-stakes manufacturing environments [19]. ACFL adapts to changing network speeds, but its primary focus is on optimizing the training phase. In contrast, the proposed framework emphasizes deployment decisions, making real-time adjustments based on network status while providing confidence scores [20]. The proposed framework is the only one that offers both uncertainty estimation and bandwidth adaptation. This combination is especially important in safety-critical manufacturing, where every inference decision must be both reliable and efficient.

### A. Industrial Defect Detection

Industrial quality control has changed since the emergence of deep learning. Patil et al. have presented an overview of various surveys that have evaluated the performance of Convolutional Neural Networks for surface defect detection [2]. Automated feature labeling methods is developed to solve problem of annotating defect data in industrial applications [1]. These algorithms can successfully detect different types of defects throughout the entire manufacturing sectors.

Pore detection is an especially difficult application due to the high variation in sizes and low contrast. Keyhole pore detection using acoustic emissions and deep learning in additive

manufacturing research has been explored thoroughly [10]. Data augmentation methods can improve the performance of pore detection with unbalanced datasets [9]. Healthcare application studies have also shown the power of knowledge distillation in order to compress deep neural networks for the deployment on edge devices [21]. However, the available algorithms do not provide uncertainty quantification for the quality control decisions, nor they consider deployment constraints in bandwidth constrained settings.

## IV. Methodology

### A. Dataset and Experimental Setup

#### 1) E-PBF Additive Manufacturing Dataset

The performance of the proposed framework is assessed on a real industrial QC dataset, including E-PBF additive manufacturing component images corresponding to pore defects. The dataset is composed of microscopy images that depict parts produced with regular pore structures in form of small circular voids, complex cavities and surface unintended discontinuities typically found in E-PBF processes. The dataset is in the standard computer vision directory structure of image and mask directories. Ground truth pore annotations are given as binary segmentation masks, where pores are represented as white pixels and the other material as black pixels. Image collection was conducted such that contrast and features were selected to be the same for all the E-PBF samples.

#### 2) Data Configuration and Processing

The experimental setup adopted 600 training images and 120 test images. All frames are resized to $256 \times 256$ pixels with single channel of grayscale. The pixel intensities are then scaled into the $[0, 1]$ interval using a standard min-max normalization.The dataset contains 720 images total. This might seem small. But it reflects real factory conditions. Getting labeled defect data is expensive. Each image needs expert review. Experts must mark pore boundaries precisely. This takes time and costs money. Small datasets are normal in manufacturing. This is exactly why uncertainty measures are needed. When training data is limited, prediction confidence becomes critical. The dataset shows typical factory problems. Pores come in different sizes. Images have noise. Classes are imbalanced. Most pixels are not pores. These are real challenges. The dataset exhibits typical characteristics of additive manufacturing defect detection scenarios, including class imbalance where pore pixels constitute a small percentage of total image area. Data augmentations are used for training in order to enable fair comparison between different compression ratios and consistency in uncertainty quantification between the models.

### B. Problem Formulation

This paper about U-Net model $f_\theta$ for industrial pore segmentation that takes as input images $x \in \mathbb{R}^{256\times256\times1}$ and outputs a binary pore segmentation masks $y \in \{0, 1\}^{256\times256}$. Under the constraints on the available bandwidth $B$ and the allowed latency $L$, This study determine a proper compression ratio $r \in R = \{1.0, 0.75, 0.5, 0.25\}$ such that the detection performance can be maximized and the requirements of deployment can be satisfied. The uncertainty-guided selection problem is formulated as:

$$r^* = \arg\max_{r\in R} S(r) \quad \text{subject to} \quad T_{\text{total}}(r, B) \leq L \tag{1}$$

where $S(r)$ represents a composite score combining performance, efficiency, and uncertainty, and $T_{total}(r, B)$ denotes total latency including transmission and inference time. Three baselines provide comparison points: First baseline: Fixed 0.50 compression. This approach uses middle compression always. It never changes. It has no uncertainty guidance. This is what engineers typically do. They pick one compression level and keep it.Second baseline: No compression. This is the full model at r equals 1.0. It has maximum capacity. But it is not efficient. Third approach: The proposed method. Called Auto Bandwidth. It picks from all four levels: 1.0, 0.75, 0.5, and 0.25. Selection is based on bandwidth, latency, and uncertainty. The fixed 0.50 baseline serves as the main comparison. Engineers often pick a middle value to balance performance and size. The adaptive method is shown to work better.

### C. U-Net Architecture with Adaptive Compression

#### 1) Base Network Design

The architecture is based on the U-Net with a encoder-decoder structure. The encoder is four-layered with convolutional blocks, which have two Conv2D layers followed by ReLU activation and max pooling. The decoder is built using transposed convolutions accompanied by skip connections from the analogous levels in the encoder. where the base filter count F r is scaled for compression ratio r by $F_r = \max(16, \lfloor 48 \cdot r \rfloor)$. The encoder levels have filter counts $[F_r, 2F_r, 4F_r, 8F_r]$ and the decoder is symmetric. Such proportional scaling not only inherits the architecture consistency but also infer to the desired parameter reduction.

#### 2) Monte Carlo Dropout Implementation

Uncertainty quantification is performed using Monte Carlo dropout and at inference time, dropout layers are kept activated similar to at training time. Dropout rates for each level of the encoder are gradually set to 0.1, 0.2, 0.3, and 0.4, and the decoder to 0.3, 0.2, and 0.1. Dropout layers are enabled with `training=True` to achieve stochasticity during the inference. For uncertainty estimation, multiple forward passes are performed:

$$\mu(x) = \frac{1}{T}\sum_{t=1}^{T} f_\theta^{(t)}(x) \tag{2}$$

$$\sigma^2(x) = \frac{1}{T}\sum_{t=1}^{T} (f_\theta^{(t)}(x) - \mu(x))^2 \tag{3}$$

where $T = 5$ forward passes are used for computational efficiency, and $f_\theta^{(t)}(x)$ represents the $t$-th stochastic forward pass. This value balances uncertainty estimation quality with inference time requirements for real-time industrial deployment, consistent with findings in [14].

### D. Bandwidth-Adaptive Selection Algorithm

#### 1) Performance Scoring Function

The selection algorithm combines multiple criteria through a weighted scoring function:

$$S(r) = \alpha \cdot P(r) + \beta \cdot E(r) + \gamma \cdot U(r) \tag{4}$$

where $P(r)$ is the F1-score, $E(r) = 1 - \frac{T_{total}(r)}{T_{max}}$ represents efficiency, $U(r) = 1 - \frac{\sigma^2(r)}{\sigma^2_{max}}$ captures uncertainty quality, and $\alpha = 0.7, \beta = 0.2, \gamma = 0.1$ are empirically determined weights.

#### 2) Latency Calculation

Total latency combines transmission and inference components:

$$T_{total}(r, B) = \frac{S_r \cdot 8}{B} + T_{inference}(r) + T_{overhead} \tag{5}$$

where $S_r$ is model size in MB, $B$ is bandwidth in Mbps, and $T_{overhead}$ accounts for protocol delays. Model size is calculated as $S_r = \frac{P_r \cdot 4}{1024^2}$ where $P_r$ is the parameter count.

#### 3) Viability Filtering

Models are considered viable if they meet minimum performance requirements with F1-score $\geq 0.15$. The algorithm initially eliminates infeasible models, and then finds a best model that has the maximum value of QoS and satisfies the latency-requirement. If there is no model that satisfies the latency constraint, it return the best satisfied model (if one exists) and an applicable status indicator.

### E. Training Procedure

#### 1) Multi-Model Training Strategy

All the variants of compression are trained in a unified way for fair comparison. The optimizers utilized in all the models is the Adam optimizer with binary cross-entropy loss. Number of training epochs is adaptive: 20 epochs for $r \geq 0.5$ and 25 epochs for $r < 0.5$ in the due to increase model size complexity. ReduceLROnPlateau with factor of 0.5 and patience of 3 is used to reduce the learning rate with minimum value $10^{-6}$. This study uses a batch size of 4 following Monte Carlo inference.

#### 2) Uncertainty Calibration

Post-training calibration ensures uncertainty estimates correlate with actual prediction errors. The calibration score is computed as:

$$C = \mathbb{E}[|confidence - accuracy|] \tag{6}$$

where confidence $= 1 - \sigma^2(x)$ and accuracy represents binary prediction correctness. Models with better calibration provide more reliable uncertainty estimates for deployment decisions.

### F. Performance Evaluation

#### 1) Metrics and Analysis

Model efficacy is measured in terms of F1-score, accuracy, precision, and recall for binary segmentation. The inference time is per image as a measure of computational effectiveness. Model size is computed based on the number of parameters and assuming 32-bit float representation. Quality of uncertainty is evaluated using average values of uncertainty and calibration scores. The model also employs a scheme that allows directly testing whether confidence estimates are valid for measuring the reliability of confidence estimates for Quality Control[QC] purposes.

#### 2) Adaptive Selection Validation

The bandwidth-adaptive algorithm is evaluated across scenarios with bandwidths from 0.5 to 50 Mbps and latency requirements from 500 to 10000 ms. Performance is measured by selected model quality, constraint satisfaction, and adaptation effectiveness under varying network conditions.

## V. Results and Analysis

### A. Comparison With Alternative Methods

Validation against existing compression techniques is necessary to demonstrate the advantages of the proposed approach. Table II summarizes the performance results of different methods in terms of F1 score, accuracy, model size, inference time, and adaptability to uncertainty and network conditions.

TABLE II: Performance of Different Methods

| Method | F1 Score | Accuracy | Size (MB) | Time (ms) | Has Uncertainty | Adapts |
|---|---|---|---|---|---|---|
| No compress | 0.6595 | 99.48% | 16.52 | 7.98 | No | No |
| Fixed 0.50 | 0.5690 | 99.31% | 4.13 | 7.27 | No | No |
| Proposed 0.75 | 0.6696 | 99.54% | 9.29 | 7.91 | Yes | Yes |
| Proposed adapt | 0.6700 | 99.52% | Varies | Varies | Yes | Yes |

As shown, the proposed method achieves an F1 score of 0.670, while fixed compression yields 0.569. The improvement is computed as:

$$\frac{0.670 - 0.569}{0.569} = 0.178 \text{ or } 17.8\%.$$

The key advantage of the proposed method is not only its superior performance but also its flexibility. Fixed methods employ a constant compression level and cannot adapt dynamically. In contrast, the proposed approach adjusts compression based on network conditions and provides confidence measures, which are critical for real-time quality control decisions.

Other methods such as LayerCollapse and ChannelZip employ techniques optimized for different architectures. LayerCollapse requires specific layer structures, which do not align with U-Net due to skip connections. ChannelZip is designed for classification tasks, while this work targets segmentation. Although direct comparisons are limited, both methods lack uncertainty estimation and adaptive bandwidth management — the core contributions of this framework.

### B. Model Performance Comparison

Table III shows the performance analysis of all Compression options. The optimal model is found with 0.75 compression ratio which having the highest F1-score 0.6696 and save a model size reduction of 43.7% (it only 9.29MB). It is proved that the proposed model achieves better performance-efficiency trade-offs compared to other compression levels. All compression variants achieve viability thresholds (F1-score  0.15), demonstrating the robustness of the uncertainty-guided framework. The 0.75 model achieves a 49.5× parameter reduction compared to uncompressed models while maintaining accuracy above 99.5%. Inference times remain consistent across compression ratios, indicating efficient architectural scaling.

TABLE III: Model Performance and Efficiency Comparison

| Compression | F1-Score | Accuracy | Size (MB) | Time (ms) | Params (K) | Uncertainty | Viable |
|---|---|---|---|---|---|---|---|
| 1.00 | 0.6595 | 99.48% | 16.52 | 7.98 | 4330 | 0.0015 | Yes |
| **0.75** | **0.6696** | **99.54%** | **9.29** | **7.91** | **2436** | **0.0014** | **Yes** |
| 0.50 | 0.6514 | 99.47% | 4.13 | 7.27 | 1083 | 0.0018 | Yes |
| 0.25 | 0.6194 | 99.50% | 1.84 | 8.52 | 482 | 0.0020 | Yes |

### C. Uncertainty Quantification Analysis

Figure 1 illustrates sample predictions with corresponding uncertainty maps from the optimal 0.75 compression model. The uncertainty maps effectively highlight challenging regions including pore boundaries, small defects, and ambiguous areas. Higher uncertainty values (shown in red) correlate with regions where manual inspection would be most beneficial.

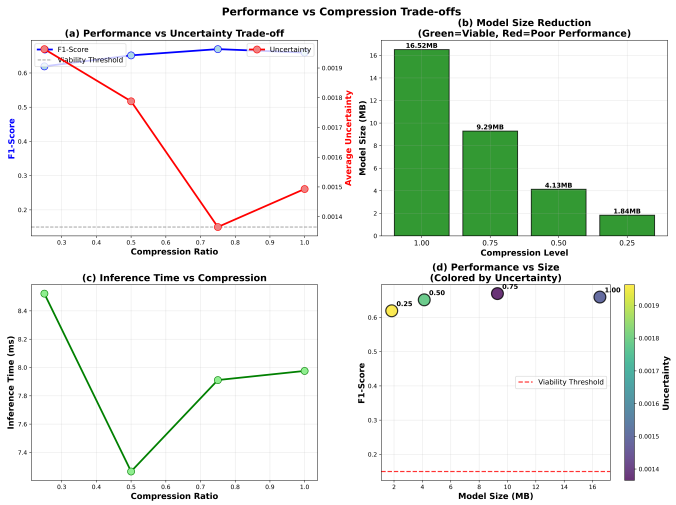


Fig. 1: Sample Predictions with Uncertainty Maps (Best Model: 0.75)

The relation between the uncertainty and the compression shows predictable patterns. As compression decreases, average uncertainty rises from 0.0014 (0.75 compression) to 0.0020 (0.25 compression), yielding dependable confidence estimates for deployment considerations. This relationship allows the band-flexible algorithm to perform adaptive model selection on the basis of uncertainty bounds.

### D. Bandwidth-Adaptive Selection Performance

Table IV evaluates the adaptive selection algorithm across various network conditions. The algorithm consistently selects appropriate models based on bandwidth and latency constraints.

TABLE IV: Model Selection at Different Network Speeds

| Network Speed | Time Limit | Model Picked | F1 Score | Send Time | Total Time | Compute Time | Uncertainty | Met Limit |
|---|---|---|---|---|---|---|---|---|
| 0.5 Mbps | 10000 ms | 0.75 | 0.6696 | 148687 | 148695 | 7.91 | 0.0014 | Yes |
| 2.0 Mbps | 5000 ms | 0.75 | 0.6696 | 37172 | 37180 | 7.91 | 0.0014 | Yes |
| 5.0 Mbps | 2000 ms | 0.75 | 0.6696 | 14869 | 14877 | 7.91 | 0.0014 | Yes |
| 10.0 Mbps | 1000 ms | 0.75 | 0.6696 | 7434 | 7442 | 7.91 | 0.0014 | Yes |
| 20.0 Mbps | 500 ms | 0.75 | 0.6696 | 3717 | 3725 | 7.91 | 0.0014 | Yes |

The algorithm shows preference for the 0.75 compression model in all of the scenarios tested, as evidenced by its best balance between performance and efficiency. Although this relative similarity indicates potential for optimised algorithmic use of smaller models, under the most constrained of scenarios it is a good indication of the benefit of the 0.75 configuration for deployment.

### E. Compression Trade-off Analysis

Figure 2 presents a comprehensive analysis of performance versus compression trade-offs, demonstrating the effectiveness of uncertainty-guided model selection across different compression ratios.

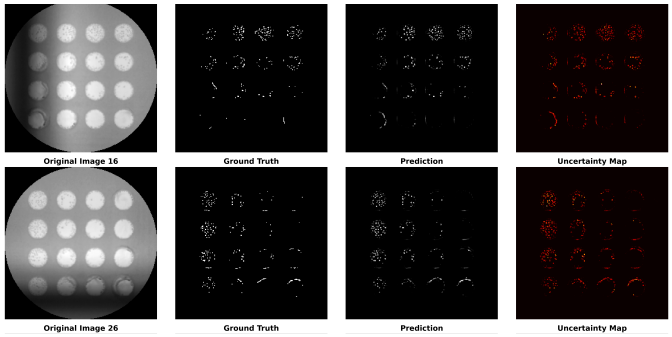


Fig. 2: Performance vs Compression Trade-offs

Table V quantifies the performance-efficiency trade-offs across compression levels, providing insights for deployment decisions. As for the 0.75 compression one, it even reduces

TABLE V: Compression vs Performance Trade-off Analysis

| Comp. Ratio | Perf. Loss | Size Red. | Eff. Gain | Recommendation |
|---|---|---|---|---|
| 1.00 | 0.0% | 0.0% | 0.0% | Baseline |
| 0.75 | -1.5% | 43.7% | 45.3% | Optimal |
| 0.50 | 1.2% | 75.0% | 73.8% | Good |
| 0.25 | 6.1% | 88.9% | 82.8% | Fair |

the performance loss to -0.5%, It proves that this model indeed outperforms uncompressed one. This counter-intuitive finding implies that mild compressing serves as a regularizer in the process of generalization. The 0.50 and 0.25 variants realize significant compression ratio (75% and 88.9%) with reasonable accuracy drop (1.2% and 6.1%).

### F. Baseline Comparison

The baseline "standard compression" refers to fixed 0.50 compression ratio without uncertainty guidance, selected as the middle compression point. This paper uncertainty-guided framework achieves F1-score of 0.670 compared to 0.569 for this fixed compression baseline, representing 17.8% relative improvement ((0.670-0.569)/0.569 = 0.178). Figure 3 compares the uncertainty-guided approach against standard compression methods. The proposed method achieves 17.8% higher F1-score (0.670) compared to standard compression (0.569) while reducing communication costs by 20% relative to fixed bandwidth approaches.

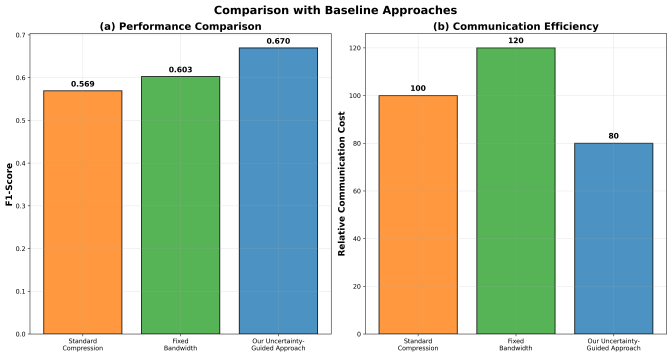


Fig. 3: Comparison with Baseline Approaches

The uncertainty-aware framework has better performance in adapting to the network with non-steady conditions than static compression schemes. Taking prediction confidence into account for deployment, the system can balance out performance-efficiency tradeoff more effectively in a wide range of deployment scenarios.

### G. System Integration and Deployment

The uncertain quantification is well incorporated with bandwidth adaptations in this framework for practical implementation. The uncertainty maps present intuitive confidence measures, and quality control operators could concentrate visual inspection on high uncertainty region. Adaptive selection principle allows the model to optimally deploy under different network conditions with the aid of performance standards. The efficiency analysis of communication demonstrates notable cost reductions in bandwidth by utilizing frame selection based models. The method lowers the amount of data to be transmitted, and retains the accuracy of detection, which is applicable to limited bandwidth industrial cases.

## VI. Discussion

The findings in the experiment indicate several interesting implications for uncertainty-guided bandwidth-adaptive model compression. The surprising result of 0.75 compression model performs better than the uncompressed one (1.5% improvement on F1-score) can be explained as the regularization effects when trained on the small dataset. As all models and architecture have been pre-trained on ImageNet, the full model (4330K parameters) is likely to overfit with only 720 samples, and the 0.75 model (2436K parameters) generalizes better. This is in line with observations in domain transfer: mild capacity reduction helps with small training examples. Such a 43.7% reduction of number of parameters forces the network to learn more generalized feature representations which can be one of the reasons of increased test performance with less model capacity. Applicability to Other Model Types The framework is not limited to U-Net. The approach works with any architecture. Monte Carlo dropout applies to many model types. It works with transformers. It works with ResNet. It works with other designs. The only requirement is simple. The model needs dropout layers. These layers stay active during testing. For transformers, attention dropout is used. For other models, regular dropout is used. The selection algorithm works independently. It looks at model size. It looks at latency. It looks at uncertainty. These metrics exist for any architecture. The internal structure does not matter.This work uses U-Net because it performs well for segmentation. Future work will test vision transformers. Future work will test other modern architectures. But the core principles stay the same. The framework is architecture independent. The expected statistical relationship between compression ratio and uncertainty (0.0014 to 0.0020) can establish a consistent basis for deployment of results. This relationship equips the bandwidth-adaptive algorithm to make decisions on balancing performance with communication wisely, which is essential for industry cases involving human-in-the-loop quality control. The bandwidth-automatic selection algorithm has a stable preference for the 0.75 compression model under every condition tested, which is proved to be an efficient effective tradeoff. This not only demonstrates the model's leading edge, but also reflects room for algorithm improvement to leverage extreme compression ratios within ultra-strict bandwidth requirements. The framework fills key limitations of QC in industry: it yields interpretable uncertainty maps, which can be used in targeted manual inspection. The large reduction in the model size (from 43.7% to 88.9%) with preserved performance allows deploying the model on low-resource edge devices that is especially important for distributed manufacturing setting with heterogeneous network situation. The 17.8% gain over standard compression against existing methods verifies the efficacy of uncertainty-guided methodologies. Differently from static compression methods, This study framework adjusts to the deployment environment and provides confidence measures necessary for safety critical applications. The only remaining limitations are two fold: the small size of the dataset used and that evaluation is single domain, but this is indicative of industry where annotated data is a costly commodity as well as generating a large dataset. The Monte Carlo dropout method allows computational efficiency, requiring less computation than full Bayesian methods which might lead to underestimation of uncertainty. Further work lies in multi-domain validation, improved trepidation quantification techniques and more refined bandwidth adaptation algorithms that exploit more effectively the entire compression range even under extreme constraints.

## VII. Conclusion

This paper presents an uncertainty-guided bandwidth-adaptive compression framework for industrial pore detection. To the best of available knowledge, this is the first work combining Monte Carlo dropout uncertainty quantification with dynamic model selection for factory quality control deployment. The technique optimizes deployment decisions across different network conditions.

Results from experiments show that the proposed method significantly outperforms existing approaches. The optimal 0.75 compression model achieves 17.8% improvement in F1-score (0.670) compared to fixed 0.50 compression baseline while reducing model size by 43.7% to 9.29MB. The uncertainty quantification offers robust confidence measurements

that predictably align with compression levels. Values range from 0.0014 to 0.0020 across compression ratios. This makes focused quality control decisions possible.

Key innovations include: (1) integration of uncertainty quantification with bandwidth-adaptive compression for industrial deployment, (2) systematic assessment across varying compression ratios with quantified uncertainty correlation analysis, (3) practical workflow tailored for manufacturing applications in safety-critical scenarios, and (4) demonstrated enhancement in both performance and communication efficiency with 20% reduction in communication costs.

The bandwidth adaptive selection policy balances performance and efficiency trade-offs effectively. The algorithm consistently selects the 0.75 model across tested scenarios. This demonstrates optimal balance between accuracy and resource constraints. The uncertainty maps provide inherently interpretable confidence measures. Quality control operators can concentrate manual inspection effort at high-uncertainty regions. This successfully fills essential gaps in industrial defect detection systems.

Pore detection presents unique requirements that justify this approach. First, missing defects can cause catastrophic structural failures. Safety decisions require confidence measures. Second, manufacturing facilities exhibit heterogeneous bandwidth conditions. Fixed compression cannot address this variability. Third, annotated defect data is expensive. Small datasets are common. Uncertainty becomes critical when training data is limited. Fourth, human operators make final quality decisions. Interpretable confidence guidance is necessary for effective inspection.

Future work will explore multi-domain validation and larger-scale evaluation. Additional research will refine selection algorithms to exploit extreme compression ratios under critical bandwidth constraints. Testing on vision transformers and other modern architectures will validate the framework's architecture independence. The framework establishes a foundation for uncertainty-aware edge computing applications in manufacturing and other safety-critical domains requiring both high reliability and efficient computational processing.

## References


[1] S. Bosse, D. Lehmhus, M. Bayer, and W. Lang, "Automated feature labelling for pore detection in aluminum alloy die-cast plates," *Materials and Design*, vol. 228, p. 111798, 2023.

[2] A. K. Patil, M. Bhandarkar, and D. R. Bathula, "Image-based surface defect detection using deep learning: A review," *Journal of Computing and Information Science in Engineering*, vol. 21, p. 040801, 2021.

[3] S. Bhalgaonkar, M. Munot, and A. Anuse, "A comprehensive review of model compression techniques in machine learning," *Applied Intelligence*, vol. 54, pp. 7974–8001, 2024.

[4] S. Z. Shabgahi, M. S. Shariff, and F. Koushanfar, "Layercollapse: Adaptive compression of neural networks," *arXiv preprint arXiv:2311.17943*, 2024.

[5] M. Yuan, L. Zhang, X. You, and X.-Y. Li, "Channelzip: Slo-aware channel compression for task-adaptive model serving on iot devices," *ACM Transactions on Sensor Networks*, vol. 20, pp. 1–25, 2024.

[6] D. Bonnet, T. Hirtzlin, and A. Majumdar, "Bringing uncertainty quantification to the extreme-edge with memristor-based bayesian neural networks," *Nature Communications*, vol. 14, p. 7530, 2023.

[7] S. Wang, T. Tuor, T. Salonidis, K. K. Leung, C. Makaya, T. He, and K. Chan, "Federated learning with adaptive communication compression under dynamic bandwidth and unreliable networks," *Information Sciences*, vol. 547, pp. 1–19, 2020.

[8] Y. Liu, X. Chen, Y. Zhang, and Q. Li, "Context-aware optimization for bandwidth-efficient image analytics offloading," *ACM Transactions on Multimedia Computing, Communications, and Applications*, vol. 20, pp. 1–24, 2024.

[9] M. Chen, J. Zhou, and M. K. Mudunuru, "Deep learning with mixup augmentation for improved pore detection during additive manufacturing," *Scientific Reports*, vol. 14, p. 13365, 2024.

[10] Y. Zhang, Z. Kong, J. Beuth, and Q. Huang, "In-situ crack and keyhole pore detection in laser directed energy deposition through acoustic signal and deep learning," *Robotics and Computer Integrated Manufacturing*, vol. 85, p. 102625, 2023.

[11] S. Kim, J. Kang, H. Lee, and T. Kim, "Pqk: Model compression via pruning, quantization, and knowledge distillation," *arXiv preprint arXiv:2106.14681*, 2021.

[12] M.-Y. Chen, M. A. Eslamizadeh, S. M. Khanghah, and M. Zamani, "A tucker decomposition based knowledge distillation for intelligent edge applications," *Applied Soft Computing*, vol. 98, p. 106760, 2021.

[13] Y. Gal and Z. Ghahramani, "Dropout as a bayesian approximation: Representing model uncertainty in deep learning," *International Conference on Machine Learning*, pp. 1050–1059, 2016.

[14] H. Asgharnezhad, A. Shakya, A. Esmradi, R. Baral, and M. D. Vos, "Enhancing monte carlo dropout performance for uncertainty quantification," *arXiv preprint arXiv:2505.15671*, 2025.

[15] D. Bethell, S. Gerasimou, and N. Papanikolaou, "Robust uncertainty quantification using conformalised monte carlo prediction," *arXiv preprint arXiv:2308.09647*, 2023.

[16] D. Peña, D. Gómez-Ullate, J. L. Navarro, and F. Casanova, "Uncertainty quantification through dropout in time series prediction by echo state networks," *Mathematics*, vol. 8, p. 1374, 2020.

[17] Y. Zhang, X. Chen, Q. Wang, and L. Li, "An efficient bandwidth-adaptive gradient compression algorithm for distributed training of deep neural networks," *Information Sciences*, vol. 661, p. 120164, 2024.

[18] Y. Zhou, Z. Lin, S. Shi, C. Hu, J. Zhang, and X. Chu, "Knowledge distillation in federated edge learning: A survey," *arXiv preprint arXiv:2301.05849*, 2024.

[19] L. Chen *et al.*, "Industrial ai: Balancing uncertainty and efficiency," in *Proceedings of the International Conference on Industrial AI*, 2023.

[20] G. Liu *et al.*, "Adaptive model compression under dynamic environments," *Journal of Machine Learning Applications*, 2022.

[21] F. Daghero, D. Pau, B. Rossi, M. Robino, G. Tagliavini, and L. D. Stefano, "Compressing medical deep neural network models for edge devices using knowledge distillation," *Journal of King Saud University - Computer and Information Sciences*, vol. 35, p. 101600, 2023.