Research Square

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

Controlled Benchmarking and Robustness Analysis of Classical Machine-Learning Approaches for Synthetic CT Sample Segmentation

Musunuri Rajani Priya
Indian Institute of Technology Roorkee
Mayank Goswami
mayankfph@iitr.ac.in

Indian Institute of Technology Roorkee https://orcid.org/0000-0001-8829-9980

Method Article

Keywords: industrial X-ray computed tomography, image segmentation, pixel-wise classification, classical machine learning, feature engineering, Decision Tree, synthetic dataset, sensitivity analysis

Posted Date: November 18th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11740273/v1

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

Additional Declarations: The authors declare no competing interests.

Abstract

Reliable delineation of a specimen from its surrounding background is a prerequisite for quantitative analysis of computed tomography (CT) images. This work is positioned as a controlled benchmarking and robustness study of classical machine-learning approaches for binary sample–background segmentation in synthetic CT-like images, rather than as a demonstration of validated industrial CT applicability. Five supervised classifiers—Logistic Regression (LR), Gaussian Naive Bayes (GNB), linear Support Vector Machine (SVM), Decision Tree (DT), and Random Forest (RF)—are compared using a synthetic dataset of 300 grayscale images (256 × 256 pixels) with corresponding binary reference masks. Three pixel-level feature representations are examined: intensity alone; intensity with normalized spatial coordinates; and a five-dimensional representation comprising intensity, two spatial coordinates, local mean, and local standard deviation calculated over a 5 × 5 neighbourhood. The dataset is divided into training, validation, and testing subsets containing 210, 45, and 45 images, respectively. Under the constructed synthetic conditions, the DT classifier using the complete feature set achieves the highest reported mean test accuracy (0.9998), Dice coefficient (0.9997), and intersection over union (IoU; 0.9995). Sensitivity experiments reveal marked performance degradation as additive Gaussian noise increases and image contrast decreases, whereas increasing the training-set size produces only marginal gains. The principal contribution is a controlled comparison of classical pixel-wise classifiers, an assessment of feature-engineering effects, and a robustness analysis under specified synthetic perturbations. The results characterize model behaviour within the defined synthetic data-generating process; they do not establish performance on experimental or industrial XCT data. Validation using independent, representative XCT scans is a necessary subsequent step.

1. Introduction

X-ray computed tomography (XCT) is a non-destructive imaging modality that reconstructs the internal attenuation distribution of an object from a set of X-ray projections. In industrial inspection and materials characterization, the reconstructed volume provides access to internal geometry and defects that may not be observable using surface-based inspection alone. Typical applications include dimensional metrology, porosity assessment, detection of inclusions, and evaluation of additively manufactured components [1, 2]. The usefulness of these measurements depends critically on the reliability of the image-processing chain, particularly the segmentation stage in which voxels or pixels are assigned to physically meaningful regions.

In a reconstructed CT image, the boundary between a specimen and its surrounding medium may be obscured by noise, limited spatial resolution, partial-volume effects, beam hardening, scatter, and other reconstruction-related artefacts. Consequently, a single global intensity threshold may not consistently separate the specimen from the background across different image regions or acquisition conditions. Classical thresholding methods, including histogram-based approaches, remain attractive because of their computational simplicity; nevertheless, their performance can deteriorate when foreground and background intensity distributions overlap [3].

Supervised machine learning offers an alternative by learning a decision function from labelled pixels and associated descriptors. In pixel-wise classification, the segmentation problem is represented as a mapping from an input feature vector—potentially including intensity, spatial position, and local neighbourhood statistics—to a binary class label. Linear and probabilistic classifiers provide relatively simple decision models, whereas tree-based methods can represent nonlinear decision boundaries and interactions among features. Random Forests extend the tree-based approach through an ensemble of decorrelated decision trees [4].

Recent studies have demonstrated the relevance of data-driven methods to industrial XCT, including segmentation and the analysis of images degraded by acquisition or reconstruction artefacts [1, 5]. For example, Jiang et al. [5] investigated semi-supervised segmentation of metal-artifact-contaminated industrial CT images, while Bellens et al. [2] compared conventional and deep-learning segmentation for XCT-based porosity measurements under reduced image quality. These studies highlight the practical importance of evaluating segmentation algorithms under realistic image-quality variations.

Against this background, the present work is explicitly framed as a controlled benchmarking and robustness study of classical machine-learning approaches for synthetic CT sample segmentation. Five classifiers are compared under a common pixel-wise classification protocol, and three feature configurations are examined to isolate the contribution of intensity, spatial position, and local neighbourhood statistics. The selected classifier is then evaluated under controlled perturbations in noise and contrast, as well as changes in training-set size. The study's contribution is methodological and comparative: it establishes a reproducible baseline for understanding how these classical models behave within a specified synthetic image-generation framework. It does not claim to demonstrate validated industrial CT applicability or to establish readiness for deployment on real XCT scans.

2. Problem Formulation and Objectives

2.1 Problem formulation

Let a grayscale image be represented by I(x,y), where x and y denote the pixel coordinates and I is the normalized intensity. The objective is to assign each pixel a binary label $y(x,y) \in \{0,1\}$, where 1 denotes the specimen (foreground) and 0 denotes the surrounding background. For a feature vector f(x,y), a trained classifier estimates the class label $\hat{y}(x,y) = h(f(x,y))$, where $h(\cdot)$ is the learned classification function. Applying this function to all pixels yields a binary segmentation mask.

2.2 Objectives

1. Generate a synthetic dataset of grayscale CT-like images with corresponding binary reference masks.
2. Compare LR, GNB, linear SVM, DT, and RF classifiers using consistent pixel-wise classification experiments.

3. Quantify the effect of intensity, spatial-coordinate, and local-statistical features on segmentation performance.
4. Evaluate the classifiers using accuracy, Dice coefficient, IoU, sensitivity, and specificity.
5. Investigate the sensitivity of the selected classifier to additive Gaussian noise, contrast reduction, and training-set size.

2.3 Scope

The present investigation is restricted to binary delineation of a synthetic specimen region from its background. Although reference CT images were consulted during the initial project stage to understand general image characteristics, the final model training and quantitative evaluation described here were performed using the synthetic dataset. The study does not address multi-material segmentation, pore or inclusion detection, three-dimensional surface reconstruction, or validation on real industrial CT scans.

Study positioning. The work is intended as a controlled benchmark rather than an application-validation study. Its conclusions are limited to relative classifier behaviour, feature-configuration effects, and sensitivity to the synthetic perturbations defined in the Methods section. Claims about practical industrial XCT performance are deliberately deferred until external validation on experimentally acquired scans is completed.

2.4.1 Study contribution and research questions

The study addresses three focused questions: (i) how do the five classical classifiers compare when trained and evaluated on the same synthetic pixel-wise segmentation task; (ii) to what extent do spatial coordinates and local-statistical descriptors improve performance relative to intensity-only features; and (iii) how sensitive is the selected classifier to controlled changes in noise level, image contrast, and training-set size? The answers are interpreted within the synthetic benchmark and are not extrapolated to real industrial XCT data.

3. Theoretical Background

3.1 Pixel-wise supervised segmentation

Pixel-wise segmentation treats each image pixel as an individual observation. The feature representation determines which image properties are available to the classifier. Intensity-only classification uses the local attenuation-related grayscale value, whereas the addition of spatial coordinates introduces positional information. Local-statistical descriptors encode neighbourhood context and may help distinguish regions with similar individual pixel intensities but different surrounding patterns.

3.2 Classifiers investigated

Logistic Regression estimates the posterior probability of a binary class using a logistic link function and, in its standard form, defines a linear decision boundary in feature space. Gaussian Naive Bayes applies Bayes' theorem under the assumptions that each feature follows a class-conditional Gaussian distribution and that features are conditionally independent given the class. A linear SVM identifies a separating hyperplane by maximizing the margin between classes. A Decision Tree recursively partitions the feature space using feature-dependent decision rules, allowing nonlinear and interaction-based classification. Random Forest combines an ensemble of decision trees, typically trained using bootstrap samples and randomized feature selection, to improve predictive stability relative to an individual tree [4].

3.3 Feature engineering

Three feature configurations were examined. Configuration F1 comprised normalized pixel intensity I(x,y). Configuration F2 comprised intensity and normalized spatial coordinates, [I(x,y), x□, y□]. Configuration F3 comprised five descriptors, [I(x,y), x□, y□, μW(x,y), σW(x,y)], where x□ and y□ are normalized image coordinates, and μW and σW denote the local mean and local standard deviation within a 5 × 5 neighbourhood W.

For a neighbourhood W containing N pixels, the local mean and population standard deviation are defined as:

$$\mu W(x,y) = (1/N) \Sigma(i,j) \in W \, I(i,j)$$

$$\sigma W(x,y) = \sqrt{\{(1/N) \Sigma(i,j) \in W \, [I(i,j) - \mu W(x,y)]^2\}}$$

The local mean describes the average neighbourhood intensity, whereas the local standard deviation quantifies local intensity dispersion. These descriptors can encode local contrast and texture-like information. Their usefulness is dependent on the scale of the neighbourhood and the spatial characteristics of the image; the 5 × 5 window used here represents the configuration evaluated in this study and should not be assumed to be universally optimal.

3.4 Performance metrics

Segmentation performance was quantified from the pixel-level confusion matrix. True positives (TP) are foreground pixels correctly labelled as foreground; true negatives (TN) are background pixels correctly labelled as background; false positives (FP) are background pixels incorrectly labelled as foreground; and false negatives (FN) are foreground pixels incorrectly labelled as background.

$$Accuracy = (TP + TN)/(TP + TN + FP + FN)$$

$$Dice = 2TP/(2TP + FP + FN)$$

$$IoU = TP/(TP + FP + FN)$$

$$Sensitivity = TP/(TP + FN)$$

Specificity = TN/(TN + FP)

Accuracy measures the fraction of all correctly classified pixels. Dice and IoU quantify foreground-region overlap, with IoU generally imposing a stricter overlap penalty. Sensitivity measures the fraction of reference foreground pixels recovered, while specificity measures the fraction of reference background pixels correctly rejected. Dice and IoU are related by Dice = 2IoU/(1 + IoU) when calculated from the same aggregated confusion counts. Because metric aggregation can be performed per image or globally, the aggregation procedure should be stated explicitly; averaging image-wise scores need not produce exact algebraic correspondence between the reported mean Dice and mean IoU [6, 7].

No single metric fully characterizes segmentation quality. In particular, accuracy may be dominated by the more prevalent class, while overlap metrics do not directly quantify boundary displacement. For practical XCT metrology, boundary-distance measures and dimensional errors may therefore be required in addition to the metrics reported here [6, 7].

4. Related Work

Machine learning has become an increasingly important component of industrial XCT workflows, spanning reconstruction, image enhancement, segmentation, surface determination, and feature characterization. Bellens et al. [1] reviewed these developments and emphasized both the opportunities and the remaining challenges associated with data-driven XCT analysis, including model explainability and uncertainty quantification. Their review provides the broader industrial context for investigating automated segmentation methods.

The difficulty of segmenting experimental XCT data is strongly influenced by image formation and reconstruction effects. Beam hardening, scattering, statistical noise, and partial-volume effects can distort local intensity distributions and make threshold-based separation unreliable. Jiang et al. [5] addressed this problem through an improved CycleGAN-based semi-supervised framework for metal-artifact-contaminated industrial CT images. Their reported results demonstrate the potential of learning-based approaches under challenging image conditions, while also illustrating the greater complexity of methods developed for artifact-affected real data.

Bellens et al. [2] evaluated conventional and deep-learning approaches for segmentation of XCT scans acquired under reduced-quality conditions for porosity measurement in polymer laser-sintered components. Their analysis considered image-quality changes associated with acquisition settings and evaluated segmentation through overlap and application-related measures. This work is particularly relevant because it links image degradation to downstream quantitative characterization rather than considering segmentation scores in isolation.

The present study differs from these investigations in its purpose and evidence base. It is not intended to compete with, or claim equivalent applicability to, methods validated on experimental industrial XCT images. Instead, it provides a controlled benchmark of five classical classifiers on a synthetic CT-like

dataset, focusing on relative classifier performance, feature engineering, and sensitivity to defined perturbations. This complementary, deliberately narrower scope makes it possible to examine algorithmic behaviour while retaining exact reference masks and controlled image conditions. Its findings should be viewed as baseline evidence for subsequent evaluation on real XCT data, not as proof of practical industrial segmentation performance.

5. Materials and Methods

5.1 Synthetic image dataset

A synthetic dataset was generated in Python to enable controlled experiments with known reference segmentations. It comprised 300 grayscale CT-like images, each with a spatial resolution of 256 × 256 pixels, and an associated binary mask. The image generator used parameterized geometric structures, including ellipses, rings, and central regions, with variations in shape, intensity, noise, and blur. The corresponding binary mask identified the prescribed foreground specimen region.

The dataset was partitioned at the image level into 210 training images (70%), 45 validation images (15%), and 45 testing images (15%). A fixed random seed of 42 was used for the split. The training subset was used to fit the classifiers, the validation subset was reserved for validation, and the testing subset was used for the reported comparative evaluation. The validation-set role and whether it was used for hyperparameter selection should be clarified in the final experimental record.

5.2 Preprocessing and feature extraction

Images were represented in grayscale, and intensity values were normalized to [0,1] during feature extraction. The three feature configurations were F1: intensity only; F2: intensity plus normalized horizontal and vertical coordinates; and F3: intensity, normalized coordinates, local mean, and local standard deviation. The two local descriptors were computed over a 5 × 5 pixel neighbourhood. This design allowed the incremental contribution of positional and neighbourhood information to be examined while retaining a common pixel-wise classification framework.

5.3 Training protocol

The five classifiers were implemented using the Scikit-learn Python library. For each training image, 1,000 foreground pixels and 1,000 background pixels were sampled, yielding 2,000 sampled observations per image and a nominally balanced training sample. This sampling strategy reduces the influence of the foreground-to-background pixel ratio during model fitting. The classifiers were trained separately for each feature configuration. Class balancing was enabled for LR, linear SVM, DT, and RF, as reported in the original experimental protocol.

The trained classifier was applied to the pixel feature vectors of each test image to produce a binary prediction mask. The DT classifier was selected for the sensitivity experiments because it achieved the highest reported mean performance with the complete feature configuration. Hyperparameters, model-

selection criteria, sampling-with/without-replacement details, software versions, and exact classifier settings were not specified in the supplied manuscript and must be added from the original code or experiment logs; they have not been inferred in this revision.

5.4 Evaluation protocol

For each test image, the predicted binary mask was compared with its corresponding reference mask. Accuracy, Dice, IoU, sensitivity, and specificity were computed at the image level, after which the mean and standard deviation across the 45 test images were determined. The standard deviation was used to describe between-image variability within the synthetic test set. The precise implementation of averaging (macro-average of per-image scores versus pooled pixel counts) should be verified and stated consistently, particularly when comparing Dice and IoU.

5.5 Sensitivity experiments

5.5.1 Gaussian-noise sensitivity

Additive Gaussian noise was introduced into the test images at nominal levels of 0, 0.05, 0.10, 0.15, and 0.20. The DT model was evaluated at each level to quantify the effect of increasing intensity perturbation on the predicted specimen boundary and foreground overlap. Fixed random seeds were used to support repeatability. The manuscript should specify whether the stated noise level represents the standard deviation of zero-mean Gaussian noise in the normalized intensity domain and whether noise was clipped to [0,1]; these details materially affect reproducibility.

5.5.2 Contrast sensitivity

Contrast was modified using factors of 1.0, 0.8, 0.6, 0.4, and 0.2 by scaling deviations from the image mean. Conceptually, this operation can be written as $I'(x,y) = m + \alpha[I(x,y) - m]$, where m is the image mean and α is the contrast factor. Values of α below unity compress the intensity distribution around the mean. The resulting images were evaluated using the DT model to assess the effect of reduced foreground–background intensity separation.

5.5.3 Training-set-size sensitivity

The DT classifier was retrained using nested subsets containing 25%, 50%, 75%, and 100% of the original training images, corresponding to 52, 105, 157, and 210 images, respectively. The subsets were obtained from a fixed random permutation, and all models were evaluated on the same 45-image test set. This nested design makes the training-set-size comparison more controlled because each larger subset contains the images used in the preceding smaller subset.

6. Results and Discussion

6.1 Comparative performance of the classifiers

Table 1 summarizes the performance of the five classifiers using the complete five-feature representation. The DT classifier achieved the highest reported mean accuracy (0.9998), Dice coefficient (0.9997), and IoU (0.9995). RF produced closely comparable values, while linear SVM also performed strongly. GNB yielded the lowest accuracy and overlap scores among the evaluated models.

Table 1
Performance comparison of classifiers using the complete five-feature representation.

Classifier	Accuracy	Dice	IoU	Sensitivity	Specificity
Logistic Regression	0.9979	0.9969	0.9938	0.9989	0.9975
Gaussian Naive Bayes	0.9788	0.9683	0.9387	1.0000	0.9686
Linear SVM	0.9989	0.9984	0.9967	0.9995	0.9986
Decision Tree	0.9998	0.9997	0.9995	0.9999	0.9998
Random Forest	0.9997	0.9996	0.9992	0.9999	0.9996

The near-unity sensitivity of GNB (1.0000) accompanied by a lower specificity (0.9686) indicates a prediction tendency favouring the foreground class: almost all reference foreground pixels were recovered, but a larger fraction of background pixels was incorrectly labelled as specimen. This behaviour is reflected in its lower Dice and IoU values. In contrast, DT and RF achieved high sensitivity and specificity simultaneously, indicating balanced classification of the two regions in this synthetic test set.

The reported standard deviations for DT were 0.000245 for accuracy and 0.000375 for Dice. These small values indicate limited variation among the 45 synthetic test images under the evaluated baseline conditions. However, low dispersion within a controlled synthetic dataset should not be equated with robustness to unrepresented acquisition systems, specimen geometries, or reconstruction artefacts. The exceptionally high scores may reflect the strong and consistent relationship between the synthetic image-generation rules, spatial layout, and reference masks. Consequently, they are best interpreted as evidence that the classifiers can learn the constructed segmentation task, rather than as an estimate of expected accuracy on experimental XCT scans.

6.2 Effect of feature configuration

Table 2 isolates the effect of feature representation for the DT classifier. The intensity-only model achieved an accuracy of 0.9963 and a Dice coefficient of 0.9944. Adding normalized coordinates increased these values to 0.9969 and 0.9953, respectively. The complete feature representation produced a further increase to 0.9998 accuracy and 0.9997 Dice.

Table 2
Decision Tree performance for the three feature configurations.

Feature configuration	Accuracy	Dice	IoU
Intensity only (F1)	0.9963	0.9944	0.9889
Intensity + coordinates (F2)	0.9969	0.9953	0.9906
All five features (F3)	0.9998	0.9997	0.9995

The improvement from F1 to F2 suggests that spatial position contains useful discriminative information in the generated images. The larger gain from F2 to F3 indicates that local mean and local standard deviation provide additional neighbourhood context beyond the intensity at an individual pixel. These local descriptors may help the tree distinguish pixels that have similar intensities but occur in neighbourhoods with different average intensity or dispersion.

This result should be interpreted in light of the synthetic geometry. When object location and shape are generated according to a constrained set of rules, normalized coordinates can act as strong predictors of class membership. The observed gain therefore does not establish that spatial coordinates will be equally informative for arbitrarily positioned or differently shaped real specimens. A useful next experiment would test the same feature configurations on specimens with randomized position, scale, orientation, and shape distributions, followed by evaluation on independent XCT scans.

6.3 Sensitivity to image noise

Table 3 reports a monotonic reduction in all three overlap/classification measures as the nominal Gaussian-noise level increases. Accuracy declined from 0.9998 at zero added noise to 0.7110 at a noise level of 0.20. Dice decreased from 0.9997 to 0.6042, while IoU fell from 0.9995 to 0.4356.

Table 3
Effect of additive Gaussian noise on Decision Tree segmentation performance.

Noise level	Accuracy	Dice	IoU
0.00	0.9998	0.9997	0.9995
0.05	0.9815	0.9720	0.9459
0.10	0.8809	0.8311	0.7151
0.15	0.7844	0.7015	0.5446
0.20	0.7110	0.6042	0.4356

The comparatively modest decline at the lowest perturbation level is followed by a sharper deterioration at higher noise levels. This behaviour is consistent with the increasing overlap of local intensity distributions and the corruption of neighbourhood statistics used by the classifier. Because F3 includes

both local mean and local standard deviation, noise can affect not only the raw intensity feature but also the contextual descriptors. The decrease in IoU is particularly pronounced, indicating a substantial loss of foreground-region overlap.

The experiment demonstrates sensitivity to the particular additive-noise model used. It does not, by itself, reproduce the full noise and artefact characteristics of industrial XCT, which may include spatially correlated noise, beam-hardening cupping, streaking, scatter, and partial-volume effects [1, 2, 5].

6.4 Sensitivity to image contrast

Table 4 shows that segmentation remained close to baseline performance at contrast factors of 1.0 and 0.8, followed by progressively larger losses as contrast was compressed. At a contrast factor of 0.2, accuracy decreased to 0.8118, Dice to 0.7722, and IoU to 0.6664.

Table 4
Effect of contrast reduction on Decision Tree segmentation performance.

Contrast factor	Accuracy	Dice	IoU
1.0	0.9998	0.9997	0.9995
0.8	0.9991	0.9986	0.9973
0.6	0.9887	0.9869	0.9770
0.4	0.9355	0.9246	0.8747
0.2	0.8118	0.7722	0.6664

Contrast compression reduces the intensity separation between the foreground and background distributions. As this separation narrows, intensity and local-statistical features become less discriminative, increasing the probability of both false-positive and false-negative classifications. The stronger degradation at low contrast is important for XCT applications because material attenuation, acquisition settings, and reconstruction conditions can influence the contrast-to-noise characteristics of the reconstructed volume [2]. The supplied results mention an increase in standard deviation at lower contrast, but numerical standard deviations are not provided in the table; these values should be added if available.

6.5 Effect of training dataset size

Table 5 presents the effect of increasing the number of training images. The DT model achieved an accuracy of 0.9995 and Dice of 0.9993 using 52 images. With all 210 training images, the corresponding values increased to 0.9998 and 0.9997. The gain is therefore small relative to the already high baseline performance.

Table 5
Effect of training-set size on Decision Tree performance.

Training fraction	Number of images	Accuracy	Dice	IoU
25%	52	0.9995	0.9993	0.9986
50%	105	0.9997	0.9995	0.9990
75%	157	0.9997	0.9996	0.9992
100%	210	0.9998	0.9997	0.9993

The diminishing improvement with additional training images suggests that the classifier learns the synthetic image-generation patterns from a relatively small subset. This may be attributable to the restricted variability and deterministic relationship between the generated shapes, their image intensities, and the ground-truth masks. The result should not be interpreted as evidence that only a small number of experimental XCT images would be sufficient for deployment. Real scans introduce broader variations in material composition, specimen geometry, scanner configuration, reconstruction parameters, and artefact structure, which are not necessarily represented by the synthetic training distribution.

6.6 Integrated interpretation

Three principal observations emerge from this controlled synthetic benchmark. First, tree-based classifiers, particularly DT and RF, achieved the strongest scores within the constructed data distribution, with DT producing the highest reported mean values. Second, adding spatial coordinates and local-statistical descriptors improved DT performance, indicating that contextual information complemented raw pixel intensity for the generated geometries. Third, the sensitivity experiments showed that baseline performance was not maintained under substantial noise or contrast degradation. In comparison, increasing the training-set size produced only marginal gains. These observations describe the behaviour of the tested classifiers under the stated synthetic data-generation and perturbation procedures; they should not be interpreted as a ranking that will necessarily persist on experimental XCT datasets.

The extremely high baseline scores warrant particular methodological scrutiny. In synthetic segmentation, a classifier may exploit regularities specific to the image generator, including a restricted family of shapes, fixed image dimensions, or consistent relationships between object position and class label. Such regularities can make the classification problem substantially easier than segmentation of experimental XCT data. A robust assessment should therefore include a deliberately shifted synthetic test distribution and, ultimately, independent real XCT scans with manually or expert-validated reference masks. This is consistent with the broader industrial XCT literature, which emphasizes the need to account for image quality, artefacts, and the intended downstream measurement task [1, 2].

A further reporting consideration is that the manuscript describes a validation subset but does not explain whether it was used for hyperparameter tuning, model selection, or any other decision. The final paper should document this explicitly. Likewise, the precise hyperparameters and implementation settings should be reported to make the comparison reproducible. Where multiple images originate from common generation templates or parameter families, the split should ensure that related variants do not appear across training and test subsets in a way that artificially inflates generalization estimates.

7. Conclusions

This work presents a controlled benchmarking and robustness study of five classical supervised-learning classifiers for binary sample–background segmentation in synthetic CT-like images. Three feature configurations were examined to quantify the contribution of pixel intensity, spatial coordinates, and local neighbourhood statistics. Within the defined synthetic benchmark, the Decision Tree classifier using the complete five-feature representation achieved the highest reported mean performance, with an accuracy of 0.9998, Dice coefficient of 0.9997, and IoU of 0.9995. Random Forest and linear SVM also achieved high scores.

The feature ablation results indicate that spatial coordinates provide a modest improvement over intensity alone, while local mean and standard deviation provide a further gain for the DT classifier. The sensitivity experiments demonstrate that performance decreases substantially under increasing additive Gaussian noise and decreasing contrast. In contrast, increasing the number of training images from 52 to 210 resulted in only a small improvement, which is likely related to the controlled nature of the synthetic dataset.

The principal outcome is a controlled benchmark of classical pixel-wise classifiers for synthetic CT sample segmentation, accompanied by an assessment of feature-engineering effects and robustness to selected image perturbations. The study provides evidence about relative model behaviour under a known synthetic data-generating process. It does not establish readiness for industrial deployment, nor does it demonstrate validated performance on real CT or XCT images. Such conclusions require independent evaluation on representative experimental scans, with documented acquisition and reconstruction conditions, reliable reference annotations, and application-relevant boundary or dimensional metrics.

8. Limitations and Future Work

The study is intentionally scoped as a synthetic-data benchmarking and robustness investigation. The findings should be considered in view of the following limitations, which define the boundary between the evidence generated here and any future claim of real-world applicability:

- Synthetic-to-real domain gap: all quantitative experiments used geometrically generated CT-like images. The dataset does not fully reproduce the attenuation physics, reconstruction artefacts, material-dependent texture, or spatially correlated noise encountered in experimental XCT.

- Restricted segmentation objective: the task was limited to binary specimen–background separation. It did not address multi-material segmentation or the identification and quantification of pores, inclusions, cracks, or other internal defects.
- Limited algorithmic scope: five classical classifiers were evaluated. Conventional thresholding and morphology-based baselines, as well as modern deep-learning methods, were not included in the reported comparison.
- Reproducibility details: classifier hyperparameters, software versions, exact synthetic-generation parameter ranges, and certain noise implementation details are not reported in the supplied manuscript.
- Generalization assessment: the same 45-image test set was used for baseline comparisons and sensitivity experiments. The study did not include an independent external dataset or a test distribution designed to assess domain shift.

Future work should evaluate the approach on experimentally acquired industrial XCT data covering multiple materials, specimen geometries, and acquisition conditions. The synthetic generator could be extended to include physically motivated attenuation distributions, spatially correlated noise, beam-hardening and streak artefacts, blur, partial-volume effects, and realistic boundary transitions. A domain-shift evaluation should reserve entire specimen geometries or generation families for testing rather than distributing closely related synthetic variants across the training and test sets.

The classifier comparison should also be expanded to include conventional thresholding (for example, Otsu thresholding [3]), morphological post-processing, and suitable deep-learning segmentation baselines. In addition to overlap measures, boundary-distance metrics and application-specific dimensional errors should be reported. Finally, uncertainty estimation, confidence calibration, and transparent reporting of model settings would strengthen the evidence required for translation into industrial XCT inspection and materials characterization [1, 6, 7].

Scope of claims. Throughout this article, reported performance values refer exclusively to the synthetic dataset and the stated test protocol. The terms “CT-like” and “synthetic CT” describe the image-generation framework and should not be read as implying that the data reproduce all physical, acquisition, or reconstruction characteristics of industrial XCT. Accordingly, the study supports comparative conclusions about the evaluated algorithms under controlled conditions, while industrial applicability remains an open validation question.

Declarations

AUTHOR VERIFICATION CHECKLIST — REMOVE BEFORE JOURNAL SUBMISSION

The following items could not be established from the supplied manuscript. They are not filled with assumptions. Please verify them against the original Python scripts, experiment logs, and dataset before submission.

Acknowledgements

The authors acknowledge the X-Ray CT Smile Facility, IIT Roorkee, for its support. The precise nature of the facility's contribution should be stated in the final submission, where applicable.

References

1. Bellens S, Guerrero P, Vandewalle P, Dewulf W (2024) Machine learning in industrial X-ray computed tomography – a review. CIRP J Manufact Sci Technol 51:324–341. https://doi.org/10.1016/j.cirpj.2024.05.004
2. Bellens S, Probst GM, Janssens M, Vandewalle P, Dewulf W (2022) Evaluating conventional and deep learning segmentation for fast X-ray CT porosity measurements of polymer laser sintered AM parts, Polymer Testing, vol. 110, art. 107540. https://doi.org/10.1016/j.polymertesting.2022.107540
3. Otsu N (1979) A threshold selection method from gray-level histograms. IEEE Trans Syst Man Cybernetics 9(1):62–66. https://doi.org/10.1109/TSMC.1979.4310076
4. Breiman L (2001) Random forests. Mach Learn 45:5–32. https://doi.org/10.1023/A:1010933404324
5. Jiang SB, Sun YW, Xu S, Zhang HX, Wu ZF (2024) Semi-supervised segmentation of metal-artifact contaminated industrial CT images using improved CycleGAN. J X-Ray Sci Technol 32(2):271–283. https://doi.org/10.3233/XST-230233
6. Taha MA, Hanbury A (2015) Metrics for evaluating 3D medical image segmentation: analysis, selection, and tool, BMC Medical Imaging, vol. 15, art. 29. https://doi.org/10.1186/s12880-015-0068-x
7. Reinke A et al (2024) Understanding metric-related pitfalls in image analysis validation. Nat Methods 21:182–194. https://doi.org/10.1038/s41592-023-02150-0