Multi-Modal Pipeline Defect Localization

Mariam Manzoor* University of Calgary
Zahra Arabi Narei* University of Calgary
Henry Leung University of Calgary
Scott Miller Baker Hughes

Abstract

This study explores the use of laser and Magnetic Flux Leakage (MFL) pipeline data to develop a deep learning model for accurate detection and segmentation of pipeline defects. Unlike conventional datasets with pixel-perfect ground truth, our labels are derived from a different sensor modality, leading to misalignment and feature discrepancies between the laser and MFL data. These discrepancies introduce label noise and domain shifts, presenting significant challenges for training models that generalize effectively. Our primary contribution lies in identifying and analyzing these challenges, and demonstrating their impact on state-of-the-art models. Through this exploratory analysis, we underscore the need for robust models and methodologies to address these fundamental issues, while outlining potential directions for future research in model design and domain adaptation strategies.

1 Introduction

Pipelines are essential for transporting oil and gas, and their integrity is critical to preventing failures and ensuring safety [1]. Magnetic Flux Leakage (MFL) and laser profilometry are widely used non-destructive evaluation (NDE) techniques for identifying pipeline defects, such as cracks, corrosion, and deformation. MFL detects anomalies by analyzing disruptions in magnetic fields, making it ideal for large-scale inspections. In contrast, laser profilometry provides high-resolution geometric profiles of the external surface, capturing precise measurements of defect shapes and sizes. Integrating these complementary modalities enhances the accuracy and reliability of defect detection, contributing to safer and more efficient pipeline operations [2].

Meanwhile, the detection accuracy of pipeline defects is significantly improved through the development of CNN networks [3], integrating MFL and laser data with CNNs still presents significant challenges. The heterogeneity of these modalities introduces alignment errors, as MFL signals and laser measurements differ in resolution, sensing principles, and spatial representation. These inconsistencies result in noisy labels and mismatches between input features and ground-truth outputs, complicating model training and reducing accuracy. Additionally, the uneven distribution of defect types and the complexity of correlating features from both modalities further intensify these challenges [2], [4].

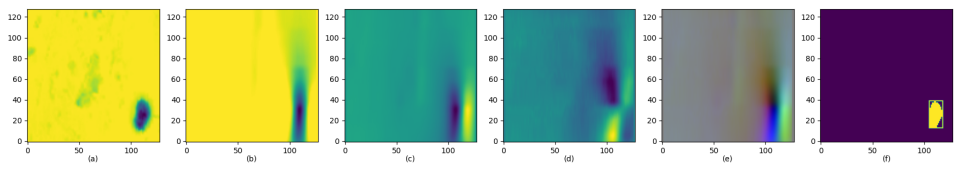

Figure 1: Labeling results: (a) Original laser signal, (b) MFL axial signal, (c) MFL radial signal, (d) MFL circumferential signal, (e) MFL 3-channel input, and (f) Defect masks bounding boxes.

*Authors contributed equally and are listed alphabetically.

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

Table 1: Detection Results for Validation Dataset

Method	Metrics(%)	Confidence Score 0.2	0.6	0.8
YOLO	AP	**50**	90	100
	AR	**50**	10	0
	F1 Score	**50**	18	0
Mask R-CNN (ResNet-18)	AP	49	65	**73**
	AR	56	51	**47**
	F1 Score	52	57	**57**
Mask R-CNN (ResNet-18-FPN)	AP	57	67	**75**
	AR	53	47	**45**
	F1 Score	55	55	**56**

Table 2: Segmentation Performance Metrics on Validation Dataset.[a]

Method	Metrics	(%)
UNet	Dice	56.5
	Jaccard	41.0
	Accuracy	77.7
Attention-UNet	Dice	65.0
	Jaccard	48.1
	Accuracy	86.4
Swin-UNETR	Dice	**72.0**
	Jaccard	**56.2**
	Accuracy	90.0
Mask R-CNN	Dice	64.3
	Jaccard	49.9
	Accuracy	**93.1**

[a] Accuracy measures the percentage of correctly predicted pixels but may overestimate performance in imbalanced datasets.

In this project, we analyze the use of the Laser and MFL Pipeline Defect dataset for training deep learning models that accurately identify and segment defects in MFL images. While Laser data provides precise defect labels, the MFL images are utilized as input features for the models. However, integrating these data sources proves challenging due to feature discrepancies between these two modalities, as deep learning models typically assume pixel-perfect labels. Through this paper, we delve into the analysis of these fundamental challenges, offering insights into their implications and proposing potential directions for developing more effective models and methodologies in the future.

2 Related Work

Pipeline defect detection using MFL data has advanced with deep learning methods. Feng et al. introduced a CNN for classifying defects as injurious or noninjurious [5], Jiang et al. proposed THMS-Net to enhance defect discrimination using heterogeneous MFL signals [6], and Yang et al. developed a multiscale SSD network with dilated convolution and attention mechanisms for small defect detection [7]. For defect size estimation, Zhang et al. introduced the VDTL network, combining radial and axial MFL signals for accurate size and profile prediction [8], while Lu et al. employed a visual transformation CNN for precise size estimation [9]. In segmentation, Wang et al. proposed Fusion PCAM R-CNN for small defect segmentation [10], and Behbahanian et al. developed PIPENet to delineate defect boundaries and reduce manual annotation [11]. Despite these advancements, existing methods struggle with label noise, boundary inconsistencies, and reliance on single-modality approaches, overlooking complementary modalities like laser profilometry. In this work, we highlight these challenges by integrating MFL images with laser data.

3 Methodology

3.1 Dataset and Labeling

In this research, we used the Laser and MFL Pipeline Defect dataset, which contains 33,000 defect samples. Each defect sample includes four signals: laser signal, MFL axial signal, MFL radial signal, and MFL circumferential signal.

The laser signal serves as the ground truth, representing the eroded percentage of the pipeline wall at each location. The input to our model is a 3-channel MFL signal image, where each channel corresponds to the axial, radial, and circumferential signals. Otsu's thresholding [12] is applied to segment the laser image into background and foreground regions, with the foreground used as the defect mask and a bounding box drawn around it. To improve generalization, we augment the data

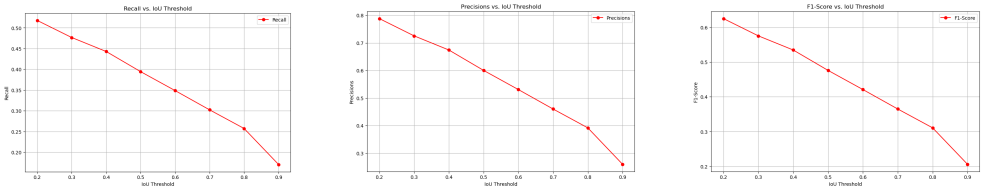

Figure 2: Recall, Precision, and F1 score vs. IoU overlap ratio on the validation set.

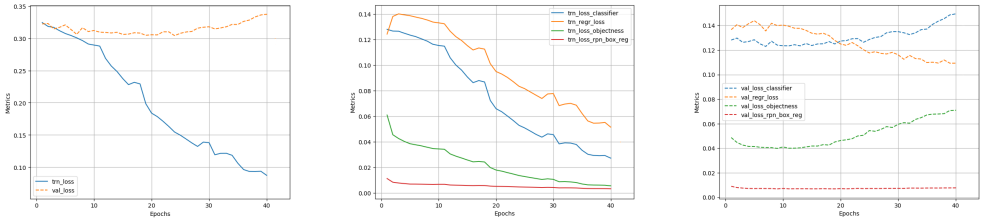

Figure 3: Training and validation losses of Mask-RCNN with ResNet-18 backbone.

using translation, zoom, and horizontal flip. Fig. 1 illustrates the defect mask and bounding box extracted from the laser image.

3.2 Detection and Segmentation Model

We employ Mask R-CNN [13] to predict defect bounding boxes and masks from 128×128 images. The model operates in two stages: first, a fully convolutional network proposes regions of interest (RoIs); then, a detection and mask head generates the bounding boxes and masks. The backbone is a ResNet-18 [14] network.

Mask R-CNN optimizes performance using multiple loss components: an objectness loss ($L_{\text{cls}}^{\text{RPN}}$) and a bounding box regression loss ($L_{\text{reg}}^{\text{RPN}}$) for the Region Proposal Network (RPN); a classification loss ($L_{\text{cls}}^{\text{head}}$), a bounding box regression loss ($L_{\text{reg}}^{\text{head}}$), and a binary cross-entropy mask loss ($L_{\text{mask}}^{\text{head}}$) for the prediction heads.

The total loss is the sum of these components:

$$\text{Total Loss} = L_{\text{cls}}^{\text{RPN}} + L_{\text{reg}}^{\text{RPN}} + L_{\text{cls}}^{\text{head}} + L_{\text{reg}}^{\text{head}} + L_{\text{mask}}^{\text{head}} \tag{1}$$

4 Results and Discussions

4.1 Defect Detection Results

We trained Mask R-CNN on our dataset. For comparison, we also trained YOLOv8 [15] and Mask R-CNN with a ResNet18-FPN backbone on our dataset.

Table 1 compares the Average Precision (AP), Average Recall (AR), and F1 score for detection models at various confidence score thresholds. The results indicate that YOLO achieves its highest AP and Recall at a very low confidence score of 0.2, demonstrating that this single-stage detector has limited confidence in its predictions. In contrast, the Mask R-CNN-based detectors exhibit higher confidence scores; however, there is little difference in performance by changing backbones, suggesting that adding a Feature Pyramid Network (FPN) [16] did not significantly enhance the results. Further details on confidence scores and the confusion matrix for these models can be found in Appendix A.

The loss curves and predictions in Fig. 3 and Fig. 4 show that the model struggles with accurate defect classification. As correct predictions that didn't align with laser-based ground truth labels were penalized, the model overfitted to noise rather than learning generalizable patterns. Additionally, as shown in Fig. 2, reducing the Intersection over Union (IoU) threshold for classifying defects improved performance, suggesting that the defect regions identified in MFL images do not always align with laser-derived ground truth.

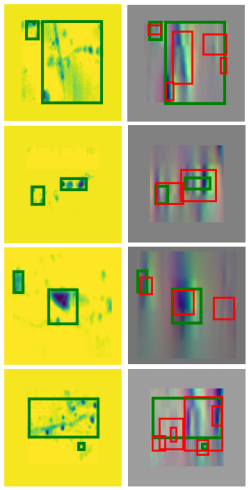

Figure 4: Defect detection results. (Left): Laser images with ground truth bounding boxes. (Right): MFL images with ground truth bounding boxes (green) and predicted bounding boxes (red).

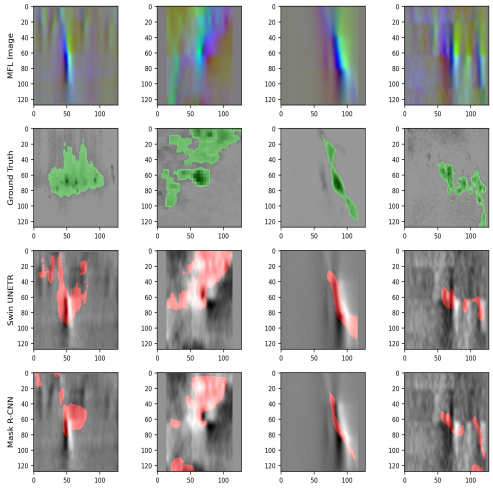

Figure 5: Segmentation results across models for various defect samples. The rows represent: (1) input MFL images, (2) laser images overlaid with the ground truth mask extracted using Otsu thresholding, (3) predicted masks by Swin-UNETR overlaid on the input MFL images, and (4) predicted masks by Mask R-CNN overlaid on the input MFL images.

4.2 Defect Segmentation Results

The segmentation experiments in Table 2 and Fig. 5 evaluate the performance of Mask R-CNN, UNet [17], Attention-UNet [18], and Swin-UNETR [19]. The results demonstrate that Mask R-CNN struggles to significantly improve Dice and Jaccard scores compared to other architectures, underscoring its limitations in accurately segmenting defects within noisy datasets. For a more detailed analysis, refer to Appendix B.

These results demonstrate interconnected challenges that impact model performance across architectures. A major issue lies in the noisy ground truth annotations derived from laser images, which often misalign with the defect features represented in the MFL images. This misalignment introduces noise into the training labels, causing the models to focus on the noise rather than capturing meaningful patterns. Furthermore, the inherent differences between laser and MFL signals create a domain gap, as MFL features lack direct point-to-point alignment with the spatial structures in laser-based ground truth masks. This issue is particularly evident in the qualitative results, where the predicted masks fail to accurately capture and align with the ground truth regions, leading to incomplete defect segmentation and lower Dice and Jaccard scores. Additionally, the shape mismatch between the ground truth masks and the actual defect regions further adds to these challenges, causing models to struggle to identify defect boundaries accurately. Despite employing different model architectures, the results indicate minimal performance differences, suggesting that the choice of model architecture is not the critical factor in improving segmentation outcomes while having cross-domain labels.

5 Conclusion

This study highlights the challenges in multi-modal pipeline defect detection, particularly the impact of noisy labels and domain shifts arising from the misalignment between laser and MFL data. By evaluating state-of-the-art models, we identify key limitations in current approaches. While this work does not propose solutions, it provides valuable insights for future research, emphasizing the need for robust models that can effectively address these challenges. Our future efforts focus on modifying loss functions, applying domain adaptation techniques, and exploring noise-robust architectures to mitigate the impact of these challenges and improve performance in detecting and segmenting defects in pipelines.

Acknowledgments

We extend our gratitude to Baker Hughes for providing the Laser and Magnetic Flux Leakage (MFL) Pipeline Defect dataset and supporting this research.

References

[1] Y. Shi, C. Zhang, R. Li, M. Cai, and G. Jia, "Theory and Application of Magnetic Flux Leakage Pipeline Detection," Sensors, vol. 15, no. 12, pp. 31036–31055, Dec. 2015, doi: https://doi.org/10.3390/s151229845.

[2] A. Belanger, D. Burden, and P. Dalfonso, "Not All Data Is Good Data: The Challenges of Using Machine Learning With ILI," Sep. 2022, doi: https://doi.org/10.1115/ipc2022-86934.

[3] S. Huang, L. Peng, H. Sun, and S. Li, "Deep Learning for Magnetic Flux Leakage Detection and Evaluation of Oil & Gas Pipelines: A Review," Energies, vol. 16, no. 3, p. 1372, Jan. 2023, doi: https://doi.org/10.3390/en16031372.

[4] S. Mukherjee, C. Hamilton, X. Huang, Lalita Udpa, and Y. Deng, "Enhanced defect detection in NDE using registration aided heterogeneous data fusion," NDT & E International, vol. 140, pp. 102964–102964, Dec. 2023, doi: https://doi.org/10.1016/j.ndteint.2023.102964.

[5] J. Feng, F. Li, S. Lu, J. Liu, and D. Ma, "Injurious or Noninjurious Defect Identification From MFL Images in Pipeline Inspection Using Convolutional Neural Network," IEEE Transactions on Instrumentation and Measurement, vol. 66, no. 7, pp. 1883–1892, Jul. 2017, doi: https://doi.org/10.1109/tim.2017.2673024.

[6] L. Jiang, H. Zhang, J. Liu, X. Shen, and H. Xu, "THMS-Net: A Two-Stage Heterogeneous Signals Mutual Supervision Network for MFL Weak Defect Detection," IEEE Transactions on Instrumentation and Measurement, vol. 71, pp. 1–9, Jan. 2022, doi: https://doi.org/10.1109/tim.2022.3198762.

[7] L. Yang, Z. Wang, and S. Gao, "Pipeline Magnetic Flux Leakage Image Detection Algorithm Based on Multiscale SSD Network," IEEE Transactions on Industrial Informatics, vol. 16, no. 1, pp. 501–509, Jan. 2020, doi: https://doi.org/10.1109/tii.2019.2926283.

[8] M. Zhang, Y. Guo, Q. Xie, Y. Zhang, D. Wang, and J. Chen, "Estimation of Defect Size and Cross-Sectional Profile for the Oil and Gas Pipeline Using Visual Deep Transfer Learning Neural Network," IEEE transactions on instrumentation and measurement, vol. 72, pp. 1–13, Jan. 2023, doi: https://doi.org/10.1109/tim.2022.3225059.

[9] S. Lu, J. Feng, H. Zhang, J. Liu, and Z. Wu, "An Estimation Method of Defect Size From MFL Image Using Visual Transformation Convolutional Neural Network," IEEE Transactions on Industrial Informatics, vol. 15, no. 1, pp. 213–224, Jan. 2019, doi: https://doi.org/10.1109/tii.2018.2828811.

[10] Z. Wang, L. Yang, T. Sun, and W. Yan, "Fusion PCAM R-CNN of Automatic Segmentation for Magnetic Flux Leakage Defects," IEEE Transactions on Neural Networks and Learning Systems, pp. 1–12, Apr. 2023, doi: https://doi.org/10.1109/tnnls.2023.3261363.

[11] Amir Behbahanian, R. Lundstrom, A. Belanger, P. Dalfonso, and R. Coleman, "PIPENet: A Semantic Segmentation Approach to Pipeline Component Detection from Magnetic Flux Leakage Readings," 2023 International Conference on Machine Learning and Applications (ICMLA), pp. 1582–1588, Dec. 2023, doi: https://doi.org/10.1109/ICMLA58977.2023.00239.

[12] N. Otsu, "A Threshold Selection Method from Gray-Level Histograms," IEEE Transactions on Systems, Man, and Cybernetics, vol. 9, no. 1, pp. 62–66, Jan. 1979, doi: https://doi.org/10.1109/tsmc.1979.4310076.

[13] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," arXiv.org, 2017. https://arxiv.org/abs/1703.06870.

[14] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," arXiv.org, Dec. 10, 2015. https://arxiv.org/abs/1512.03385.

[15] D. Reis, J. Kupec, J. Hong, and A. Daoudi, "Real-Time Flying Object Detection with YOLOv8," May 2023, doi: https://doi.org/10.48550/arxiv.2305.09972.

[16] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature Pyramid Networks for Object Detection," arXiv:1612.03144 [cs], Apr. 2017, Available: https://arxiv.org/abs/1612.03144.

[17] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," arXiv.org, May 18, 2015. https://arxiv.org/abs/1505.04597.

[18] O. Oktay et al., "Attention U-Net: Learning Where to Look for the Pancreas," arXiv.org, 2018. https://arxiv.org/abs/1804.03999.

[19] A. Hatamizadeh, V. Nath, Y. Tang, D. Yang, H. Roth, and D. Xu, "Swin UNETR: Swin Transformers for Semantic Segmentation of Brain Tumors in MRI Images," arXiv:2201.01266 [cs, eess], Jan. 2022, Available: https://arxiv.org/abs/2201.01266.

A Extended Analysis of Defect Detection Results

The validation set detection results using Mask R-CNN are analyzed in detail in this section. Fig. 6 illustrates the trends in precision, recall, and F1-score across varying confidence thresholds, showing that precision increases with higher thresholds due to reduced false positives, while recall decreases as fewer true positives are captured, reflecting the tradeoff between these metrics. Notably, recall remains below 100%, likely due to noisy supervision during evaluation, which hinders the model's ability to detect all defects accurately. Fig. 7 provides a confusion matrix, offering a breakdown of true positives, false positives, false negatives, and true negatives. The confusion matrix highlights a significant number of false positives and false negatives, indicating that label noise and domain shifts in the dataset significantly impact the model's performance, emphasizing the need for more robust detection approaches.

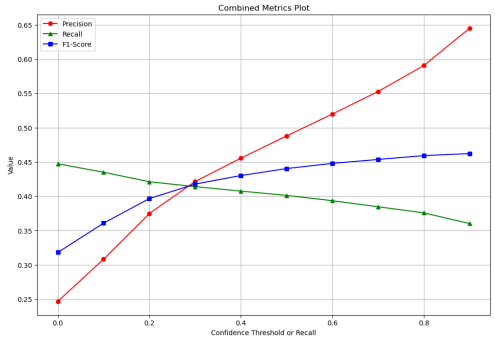

Figure 6: Precision, recall, and F1-score trends across varying confidence thresholds.

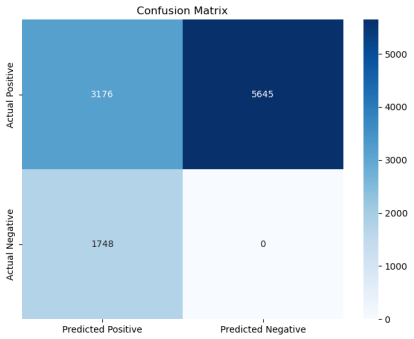

Figure 7: Confusion matrix for the validation set.

B Extended Analysis of Defect Segmentation Results

The analysis of the segmentation results provides critical insights into the impact of label noise on the model's performance. The distribution of Dice scores for the validation set, shown in Fig. 8, highlights significant variability across the dataset. While a subset of samples achieves high scores, reflecting effective segmentation, the majority fall into intermediate ranges, where predictions are uncertain or inconsistent. This middle range suggests the model struggles to align predictions with ground truth due to challenges such as label noise and misalignment. Moreover, a notable portion of samples exhibits low Dice scores, underscoring the difficulty of reliably capturing defect regions under noisy supervision.

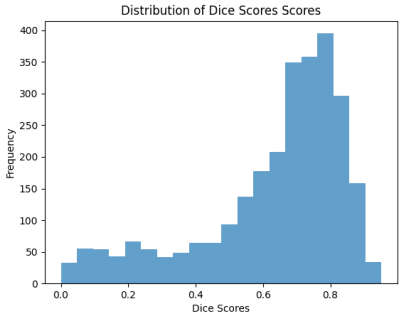

Figure 8: Distribution of Dice scores for the validation set.

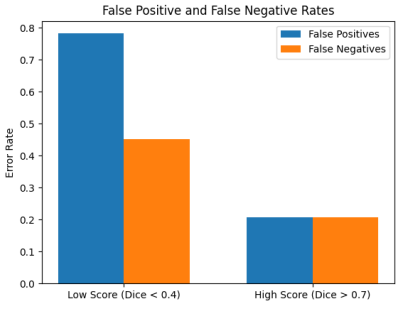

Figure 9: False positive and false negative rates for the low-scoring group with Dice scores below 0.4 and the high-scoring group with Dice scores above 0.7.

To further examine the low-scoring samples based on Dice scores, Fig. 9 analyzes false positive and false negative rates for both low- and high-scoring groups. Low-scoring samples are dominated by a

high false positive rate, indicating the model's tendency to over-segment ambiguous regions, and a moderately high false negative rate, reflecting difficulties in identifying all defect regions. In contrast, high-scoring samples demonstrate both low false positive and false negative rates, showcasing the model's ability to generalize effectively when provided with clean and well-aligned ground truth labels.

Fig. 10 complements this analysis by exploring the relationship between the defect area of each sample in the ground truth and the corresponding predicted defect area for low-scoring samples. The scatter plot reveals substantial overpredictions for smaller defects, with predicted areas frequently exceeding the ground truth. This highlights the model's challenge in dealing with noisy labels, particularly for small defect regions, which disrupts the accuracy of its predictions.

Together, these findings underscore the profound impact of label noise on model performance, manifesting as inconsistent Dice scores, a tradeoff between false positives and false negatives, and significant discrepancies in defect area predictions.

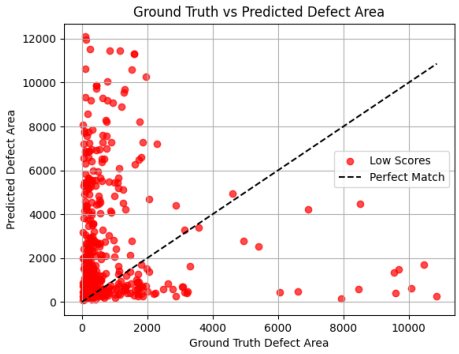

Figure 10: Comparison between the defect areas of each sample in the ground truth and the corresponding predicted defect areas.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Abstract and Introduction

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The limitations are discussed in the Results and Discussions sections

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

Justification: No theoretical assumptions or proofs are used in this work, as it is based on practical implementations.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The data used is private and not publicly available. However, the same experiment can be repeated on any other dataset as architecture used is disclosed.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [No]

Justification: The architecture used is commonly available, but the dataset is private and cannot be publicly disclosed.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [No]

   Justification: The submission for Muslim in ML is up to 4 pages and it was a challenge to include all the details within 4 pages.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Refer to Results and Discussions Section

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The submission for Muslim in ML is up to 4 pages and it was a challenge to include all the details within 4 pages.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The work respects data usage agreements, ensures proper citation and acknowledgment of external resources.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: This research focuses on improving pipeline defect detection. The potential positive societal impacts include enhanced pipeline safety, reduced environmental risks, and improved operational efficiency in the oil and gas industry.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.

- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [No]

    Justification: This paper does not release any new models, datasets, or assets that pose a high risk of misuse.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The dataset used in this research is a proprietary resource provided with proper consent and acknowledgment from the data provider.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.

- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: This paper does not introduce new assets. The research utilizes a proprietary dataset provided with appropriate consent. No new datasets, models, or code assets are released as part of this work; instead, the focus is on analysis.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: This paper does not involve crowdsourcing or research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: This paper does not involve crowdsourcing or research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.