Proceedings of ICFTBA 2026 Symposium: Driving Organizational Strategic Change for Financial Gains
DOI: 10.54254/2754-1169/2026.BH38189

Transfer Learning with ResNet-18 for Surface Defect Inspection in Hot-Rolled Steel

Michael Guan

High School Attached to Northeast Normal University, Changchun, China
2759059361@qq.com

Abstract. Surface inspection in hot-rolled steel manufacturing requires decisions about material moving through a production line. A model that distinguishes photographed defects must also handle normal surfaces, changing acquisition conditions and uncertain predictions before its output can support coil release. This paper develops a research framework for adapting ResNet-18 to this setting with limited factory annotations. The proposed pipeline combines line-scan image acquisition, patch classification and a decision policy that assigns cases to pass, human review or hold. The NEU surface defect database supplies source-domain examples, factory images supply a normal class and the target-domain evidence required for deployment. The proposed experiment compares training budgets of 10, 25 and 50 images per class, with separate coils allocated to training, validation and testing. Comparisons distinguish six-class defect recognition from seven-class factory inspection, preventing misleading comparisons between models with different label spaces. Evaluation covers class performance, defect escapes, unnecessary holds, review workload and processing deadlines. The analysis identifies the conditions under which the experiment could support deployment, particularly representative normal data and independent validation of release decisions. No model training or factory trial is reported. The contribution is an experimentally testable inspection framework connecting transfer learning to operational quality control.

Keywords: hot-rolled steel, surface inspection, transfer learning, ResNet-18, human review

1. Introduction

Surface inspection connects the condition of a steel strip to decisions about its processing and release. A missed defect can carry a quality problem into subsequent operations, while an unnecessary rejection removes usable material from production. Gao et al. place quality assurance within the wider use of artificial intelligence in manufacturing [1]. For a steel line, the practical issue is how an image-based prediction can become a traceable decision under production constraints. Luo and He describe an integrated surface inspection system for hot-rolled flat steel, illustrating the importance of acquisition and system design alongside recognition algorithms [2].

Defect appearance complicates that task. Scratches, inclusions and irregular surface textures vary within a category, and different categories can look similar. Song and Yan address noise and appearance variation using adjacent evaluation completed local binary patterns (AECLBP) [3]. A

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2754-1169/2026.BH38189

learned classifier offers another way to represent these textures, but its usefulness on one image collection does not establish its reliability under a different camera, steel grade or lighting condition.

This paper develops a method for transferring a ResNet-18 classifier to a factory line when annotations are limited. The central hypothesis is that a small, carefully sampled set of factory images will improve performance on unseen coils relative to a model trained only on the source dataset. A second requirement is that the resulting decisions remain useful at an acceptable review workload and processing speed. The study defines the data, comparisons and deployment criteria needed to test these propositions, it does not claim completed experiments.

2. Research background and experimental design

2.1. Related work and research scope

Steel surface inspection includes classification, localisation and the interpretation of defects against product requirements. These tasks require different outputs. He et al. present a detection approach that fuses hierarchical features [4]. The present framework instead classifies fixed image patches. Their recorded acquisition coordinates locate suspicious regions approximately, the classifier does not produce bounding boxes or defect boundaries. This narrower scope permits a focused experiment with image-level labels.

ResNet introduces residual connections that allow stacked layers to learn modifications to their inputs [5]. ResNet-18 is selected as a candidate whose computational demands can be measured on factory hardware. Its selection is provisional: greater recognition complexity must justify its costs against a texture-based comparator. The LBP–SVM baseline proposed here uses ordinary local binary pattern features and a support vector machine. It is not a reproduction of Song and Yan's AECLBP method [3].

Transfer learning uses knowledge from a source task or domain to support learning in a target setting. Pan and Yang distinguish differences in domains and tasks that affect transfer [6]. Both matter here: factory acquisition changes the image distribution, and adding normal material changes the label space. The proposed contribution is therefore a controlled evaluation and decision framework, rather than a new neural architecture. It tests whether adaptation helps under stated annotation budgets and whether that improvement supports production decisions.

2.2. Source data and factory sampling

The NEU database contains 1,800 grayscale images at 200 × 200 pixels, with 300 examples for each of six defect categories: rolled-in scale, patches, crazing, pitted surface, inclusions and scratches [7]. The classification collection supplies examples of known defects but no normal category. Source training alone therefore cannot establish that a factory patch is acceptable for release.

The target collection must contain representative normal surfaces and defects from the intended line. Each patch retains its coil identifier, acquisition time and position, available camera settings and steel grade are recorded. Inspectors assign labels using a written guide. Ambiguous or mixed-defect patches receive an explicit review flag instead of a forced single-class label. Such cases remain in the operational evaluation, even when excluded from the controlled seven-class classification comparison.

Normal examples must include ordinary texture variation and acceptable surface features. Defective examples should span different coils and conditions. Increasing the number of adjacent

DOI: 10.54254/2754-1169/2026.BH38189

images from one defect does not provide the same evidence as observing that category on several independent coils.

2.3. Separation and annotation budgets

Complete coils are allocated to training, validation or testing before patches are selected. Neighbouring and overlapping patches from a coil stay together. A fixed test collection is retained for all comparisons, with its labels unavailable during development. If a defect category occurs on too few independent coils, that limitation is reported and the corresponding comparison is deferred. Randomly distributing adjacent frames cannot repair missing independent evidence.

The training budgets are 10, 25 and 50 labelled target patches per class. For seven classes, these correspond to 70, 175 and 350 training labels. Validation and test annotations are additional costs and must be reported separately. Within each repeated sampling run, the smaller budgets are nested inside the larger ones. Within each task and budget, target-trained models use identical sampled target patches and identical test coils, so differences cannot be attributed to favourable sample selection.

2.4. Comparable tasks

Table 1 separates the domain-transfer diagnostic from the deployment comparison. The source-only model has six outputs and is evaluated only on the six known defect categories. For this diagnostic, adapted models use a six-class head trained without normal examples. The seven-class factory experiment compares methods trained with the same target labels, including normal material. A six-class source model is never assigned a normal prediction through an arbitrary confidence threshold.

Table 1. Planned comparisons and their evidence requirements

Evaluation	Models and target training labels	Test scope
Domain transfer	Source-only ResNet-18, six-class adapted ResNet-18 at 10, 25 and 50 labels per class	The same unseen-coil subset containing six known defects
Factory inspection	LBP–SVM, ImageNet-to-target ResNet-18, ImageNet-to-NEU-to-target ResNet-18 at equal budgets	Normal material and six defects from fixed unseen coils
Operational decisions	Frozen seven-class model and decision thresholds	The full inspection stream, including ambiguous cases

The direct ImageNet-to-target model controls for the intermediate NEU stage. It uses the same architecture and target training protocol as the proposed model, while omitting source defect training. This additional comparison distinguishes the value of NEU adaptation from the effects of ImageNet initialisation and access to factory labels.

3. Inspection method

3.1. Acquisition and model adaptation

A line-scan camera records successive strips of the moving surface. The acquisition system assembles patches and associates them with coil coordinates. Exposure, focus and coverage checks identify unusable input before classification. A failed check generates a review or hold event because an unreadable image supplies no evidence for release.

DOI: 10.54254/2754-1169/2026.BH38189

Images are resized to 224 × 224 pixels, and greyscale values are repeated across three channels for compatibility with ImageNet-initialised weights. Resizing standardises the input, it cannot recover surface detail absent from acquisition. The pretrained ResNet-18 receives a six-output classification head and is first adapted to NEU. For the factory experiment, a seven-output head is fitted using target training data, followed by selective unfreezing of the final residual stage. The earlier stages remain fixed initially to limit the number of parameters estimated from scarce labels.

Cross-entropy is the proposed training objective. Checkpoint selection uses validation macro-F1, with early stopping and a small, declared learning-rate search. Source-stage settings and checkpoints use a held-out NEU subset, target-coil validation is reserved for target-trained models and the decision policy. Candidate augmentation is limited to changes that preserve defect identity. Strong crops or rotations that remove small defects or alter meaningful orientation are excluded. The LBP–SVM comparator receives the same target patches and validation access, with feature settings and SVM parameters recorded. These choices are protocol specifications, not reported training outcomes.

3.2. Confidence and production decisions

A softmax score expresses the model's relative support for its available classes, it does not establish that an image belongs to the training distribution. Guo et al. show why neural-network confidence may require calibration and describe temperature scaling [8]. Here, temperature scaling and decision thresholds are fitted on separate validation portions grouped by coil. Calibration quality is checked before scores are interpreted as probabilities. Sparse validation evidence may justify retaining human decisions throughout the pilot.

The pass threshold applies only to the normal class after image-quality checks succeed. A high-confidence defect prediction triggers a hold for assessment. Final rejection depends on the relevant product specification and an authorised inspector, since defect category alone does not determine severity. Remaining cases enter human review. Unfamiliar images can still receive high confidence, so this policy does not claim a validated open-set detector. Monitoring acquisition conditions and auditing accepted material remain necessary.

A patch-level pass is evidence about one inspected region. Coil release requires adequate coverage and resolution of every associated review or hold. Logs retain the image, model version, score, decision and inspector outcome. Verified outcomes may inform later retraining, but the running model remains fixed until the revised version passes validation.

4. Evaluation and discussion

4.1. Recognition and operational errors

The planned evaluation uses complementary classification measures to examine different errors [9]. Macro-F1 weights each class equally, per-class recall reveals categories that are missed disproportionately. Both will be reported with class counts and a confusion matrix. Table 2 defines operational measures separately, since a correct class label does not by itself establish an appropriate release decision.

Table 2. Proposed evaluation measures and reporting units

Measure	Definition or required report
Macro-F1	Mean class F1, with the label set stated

DOI: 10.54254/2754-1169/2026.BH38189

Table 2. (continued)

Defect escape rate	Defective patches automatically passed / all defective patches
Unnecessary hold rate	Acceptable patches held / all acceptable patches
Review routing rate	Patches assigned to human review / all inspected patches
Timing	End-to-end latency distribution, throughput and missed deadlines

Class-balanced subsets support controlled recognition comparisons. Workload and escape estimates also require a naturally sampled inspection stream with its class proportions stated. Artificially balancing that stream would distort the expected frequency of reviews and holds. Inspector decisions are checked against adjudicated labels, since human review can also introduce errors.

Automatic coverage, the fraction receiving an automatic pass or hold, is reported alongside review routing. Hold cases also require assessment, review and hold counts are therefore reported separately to describe human workload. Threshold comparisons show how escape risk changes with coverage. Routing nearly every image to review may reduce automatic errors while providing little operational benefit. Holding acceptable material is distinguished from rejecting it after inspection. Coil-level release errors are reported separately from patch-level rates to avoid confusing local recognition with product disposition.

Label sampling and model fitting will be repeated using five declared random seeds. Results will include individual runs and their mean and spread. Uncertainty intervals resample whole test coils, preserving dependence among patches from the same coil. A small number of coils limits the strength of these intervals. An observed zero-escape count is accompanied by the number of defective patches and coils examined, it is not treated as proof that the true escape risk is zero.

4.2. Timing and pilot progression

Timing starts at acquisition and ends when the decision is available to the control interface. Measurements include patch assembly, preprocessing, inference, queues and communication. For a line-scan system, the interval between individual scan lines differs from the interval between completed patches. Throughput must match the patch arrival rate, while decision latency must fit the physical time available before the relevant downstream action. The p95 timing check is used as a diagnostic, supplemented by higher-percentile latency, deadline misses and queue growth. A percentile alone cannot guarantee timely processing of every patch.

Profiling records hardware, input dimensions, batch size, numerical precision and workload duration. Model storage, multiply–accumulate operations and energy per processed image describe different resource costs. Energy measurement states its system boundary and idle-power treatment. None can be inferred reliably from the architecture name alone.

Deployment begins in shadow mode, with predictions recorded while existing inspection controls release. A human-assisted pilot follows only after recognition and timing are assessed on independent coils. The NIST AI Risk Management Framework supports documenting responsibilities, monitoring and the handling of risks across system use [10]. Quality staff set application-specific error limits before the final test. A failed quality gate or persistent processing backlog returns the workflow to supervised inspection.

DOI: 10.54254/2754-1169/2026.BH38189

4.3. Interpretation and limitations

The experiment can test whether limited target annotations improve recognition, but it cannot establish that 10, 25 or 50 images per class are sufficient for every line. These budgets describe controlled comparisons. Rare defects, normal-surface diversity and changes in illumination may demand a larger collection, particularly for validation of automatic release.

A useful result would combine better defect recognition with an acceptable review burden at a stated escape limit. Improved macro-F1 accompanied by excessive holds would leave the production problem unresolved. If the texture baseline performs similarly at lower cost, retaining that method would be a valid outcome. Negative transfer would also be informative, indicating that source features or adaptation settings require reconsideration.

The present analysis has no trained models, independent factory collection or hardware measurements. Its conclusions concern experimental design and the conditions for interpreting future evidence. Claims about accuracy gains, real-time capability, reduced scrap or financial savings remain outside the available evidence.

5. Conclusion

The proposed framework connects ResNet-18 transfer learning with surface inspection decisions for hot-rolled steel. It defines a path from line-scan acquisition to patch classification and supervised handling of uncertain material. Its main methodological requirement is consistency between the task a model has learned and the decision its output is used to support. Six defect categories provide a starting point for recognition, while automatic acceptance requires representative normal material and validation under the intended factory conditions.

The experimental design addresses this requirement through separate six-class and seven-class comparisons. Whole-coil separation limits leakage from neighbouring images, and shared training budgets permit comparison under the same annotation constraints. Recording the additional cost of validation and test labels makes the proposed small-data experiment interpretable. The comparison can then show whether target adaptation provides a useful improvement and how that improvement changes as more labels become available.

Operational evaluation extends beyond recognition scores. Defect escapes, unnecessary holds and review workload describe how the system affects inspection decisions. Timing measurements must reflect patch arrival and the downstream decision deadline, with sufficient throughput to prevent a growing queue. Confidence thresholds require validation and cannot ensure that unfamiliar images will always be rejected for review.

The next empirical step is to assemble independently labelled factory coils and execute the fixed protocol. Until those observations exist, the framework supports a testable research proposal rather than a claim of successful deployment. Automated release becomes defensible only when the selected model, thresholds, acquisition coverage and control response jointly satisfy the factory's stated acceptance criteria. Human authority over unresolved cases remains part of that decision process.

References

[1] Gao, R. X., Krüger, J., Merklein, M., Möhring, H. C., & Váncza, J. (2024). Artificial intelligence in manufacturing: State of the art, perspectives, and future directions. *CIRP Annals, 73*(2), 723–749.

[2] Luo, Q., & He, Y. (2016). A cost-effective and automatic surface defect inspection system for hot-rolled flat steel. *Robotics and Computer-Integrated Manufacturing, 38*, 16–30.

DOI: 10.54254/2754-1169/2026.BH38189

[3] Song, K., & Yan, Y. (2013). A noise robust method based on completed local binary patterns for hot-rolled steel strip surface defects. *Applied Surface Science, 285*, 858–864.
[4] He, Y., Song, K., Meng, Q., & Yan, Y. (2020). An end-to-end steel surface defect detection approach via fusing multiple hierarchical features. *IEEE Transactions on Instrumentation and Measurement, 69*(4), 1493–1504.
[5] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778.
[6] Pan, S. J., & Yang, Q. (2010). A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering, 22*(10), 1345–1359.
[7] Song, K., & Yan, Y. (2026). NEU surface defect database. Northeastern University. (accessed 14 September) https://faculty.neu.edu.cn/songkechen/zh_CN/zdylm/263270/list/index.htm.
[8] Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017). On calibration of modern neural networks. *Proceedings of Machine Learning Research, 70*, 1321–1330.
[9] Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. *Information Processing & Management, 45*(4), 427–437.
[10] Tabassi, E. (2023). Artificial intelligence risk management framework (AI RMF 1.0). NIST AI 100-1. National Institute of Standards and Technology. https://doi.org/10.6028/NIST.AI.100-1