Journal of Applied Computational Intelligence

eISSN: 3136-6937

Vol. 4, No. 2, April, 2026

Transformer Model for Rail Squat Classification with Lift off Invariant Eddy Current Embeddings

Oktay Savaş[1], Mesut Rauf[1,*]

[1] *Department of Civil Engineering, Faculty of Engineering, Yıldız Technical University, Istanbul, 34220, Turkey*

Received 18 November 2025; Revised 26 March 2026; Accepted 29 March 2026; Published 05 April 2026

Abstract

Probe lift-off changes the amplitude, phase and spatial extent of eddy-current responses, complicating classification of rail-squat indication morphology. This study evaluates a two-layer Transformer with lift-off-invariant projected embeddings on a declared numerical response family. Three complex observation channels at nominal excitation settings 25, 50 and 100 kHz are sampled at 96 positions across an 80 mm window. Training-derived channel standardization, window centering and common energy normalization precede 24 ordered spatial tokens. Paired observations of one geometry receive a consistency objective, while supervised contrastive learning organizes projected vectors by four reference families: reference, compact, extended and compound. Independent geometry partitions contain 1,200 training, 320 validation and 640 held-out sources. Each held-out geometry is observed in a seen lift-off interval of 0.5-2.5 mm and an unseen interval of 2.8-4.0 mm. Under the unseen condition, the full model reaches 97.19% accuracy and macro-F1 0.9720, versus 95.00% and 0.9500 for a raw-input Transformer. Its paired-embedding cosine similarity is 0.9655, compared with 0.9226 for the raw model, and seen-range ridge-probe lift-off R squared is 0.0684 versus 0.3491. A class-only objective reaches higher unseen accuracy of 98.28% in the primary run. Validation-selected temperature 0.5 increases unseen expected calibration error from 0.0155 to 0.0241, indicating weaker probability transfer than class decision transfer. At 4.1-5.0 mm lift-off, full-model accuracy is 73.28%; added noise and central evidence gaps yield 76.56% and 74.53%. These outcomes distinguish classification, representation consistency and confidence within the synthetic condition-shift setting.

Keywords: Rail Squat Classification; Eddy Current Signals; Lift Off Invariance; Transformer; Contrastive Learning

1. Introduction

Rail surface inspection increasingly combines electromagnetic observations with computational models that organise local responses into defect-related categories [1]. Eddy current sensing provides spatially resolved information through changes in electromagnetic coupling near the rail head [2]. Probe lift-off modifies that coupling and can alter an indication even when its underlying geometry remains fixed [3]. Representation learning offers a mechanism for separating repeatable response structure from observation-dependent variation [4]. For squat classification, the computational task is to preserve information about local extent, asymmetry and adjacent indications while reducing sensitivity to the probe separation. A classifier trained on a narrow operating range can otherwise associate attenuated or broadened observations with a different category, producing condition-dependent inspection decisions.

The four numerical classes are assigned from source-response width and lobe structure before observation conditions change. Attenuation weakens response magnitude, phase rotation changes real-imaginary balance, and spatial

* Corresponding author: rauf.me@yildiz.edu.tr

http://dx.doi.org/10.68406/jaci.2026.vol4iss2no1:1-13

➢ This is an open access article under the CC BY-NC-SA 4.0 license (https://creativecommons.org/licenses/by/4.0/).

© Authors retain all copyrights.

broadening alters neighboring sample relationships. Learning under distribution shift [5] therefore requires evaluation beyond the training lift-off interval. Multiple observations of one source geometry share its morphology and class label. Grouped evaluation [6] assigns those related views to one data partition before training and scoring. Held-out geometries observed under both seen and unseen lift-off consequently provide a controlled comparison of condition change under fixed source labels.

An invariant representation must retain the morphology needed to distinguish compact, extended and compound indications while reducing lift-off-driven variation. Contrastive representation learning [7] relates observations through their source identity or shared class, and attention-based sequence processing [8] aggregates ordered spatial cues across the response window. Paired consistency aligns two conditions of one geometry; supervised contrastive learning organizes distinct geometries by category. Their contributions are evaluated separately because repeatability and category discrimination are distinct targets. Probability calibration [9] provides a third assessment of whether confidence transfers with the class decision to an unseen condition interval.

This study evaluates a compact Transformer with window normalization, paired lift-off consistency and supervised contrastive projection learning for four synthetic rail-response families. Grouped geometry partitions prevent repeated views of a source from crossing training and scoring boundaries. Section 2 reviews eddy-current signal descriptors, condition-invariant learning and attention-based classification. Section 3 defines the complex response generator, training-derived standardization, spatial tokens, shared encoder, representation losses and probability selection. Section 4 reports unseen-condition classification, embedding measurements, component ablations, disturbance sensitivity and retrospective review capture. Section 5 summarizes three quantitative findings and the requirements for measured-rail validation.

2 Related Work

2.1 Eddy Current Signal Processing for Rail Inspection

The first types of eddy current classifications are generally by amplitude, phase, differential response, etc. Computational signal analysis [10] connects these descriptors with the spatial structure of a recorded signal. A response peak is a local deviation, and changes in the neighbouring polarity and the relationship among quadrature components indicate the size and shape of the peak. A moving probe is also employed to obtain sample data on the entry and exit of an indication from the sensitive area. A SpatialSequence retains these connections and can be used by a classifier to compare narrow, broad, and overlapping patterns in the same observation window. This form is especially suitable when the individual indications have the same maximum amplitude.

Change in Conditions Modifies the Effect of Attributes. Multi-channel feature learning [11] acquires various observations, and normalisation methods [12] reduce systematic scale differences prior to classification. These changes have different information costs. Removal of a window is equivalent to a constant offset; division by the total energy of a window normalises the amplitude scale across all channels. Relative channel amplitude, phase and spatial shape can still be obtained after this operation. Signal-quality assessment [13] is also relevant, as the energy normalisation increases the relative contribution of noise in weak observations. Therefore, the two parts of the computational design are a regularised denominator and explicit low-signal evaluation.

2.2 Domain Invariant Feature Learning

Invariant learning learns representations that are still effective under different conditions of observation. Paired Representation Objectives [14] make this idea feasible by connecting two views with a known common source. View construction specifies which parts of the model should be ignored. For the current task, the two lift-off observations have the same shape and class but different attenuation, phase and noise. They are closer to being the same as siblings because they share the same parent geometry. The consistency objective aims to have the same projected embeddings for this pair, but the classifier still receives the full encoded sequence and learns its association with the reference category.

Class-conditioned contrastive objectives [15] extend the relationship to different geometries within the same class. They organise a representation by showing similarities to positive and negative observations in a batch. The larger organisation can maintain category separation by means of strict paired alignment alone. Domain-generalisation evaluation [16] then shows the representation in a held-out condition interval. It is used to report both classification and reconstruction errors. Although high paired similarity indicates stability, class separation and the predictability of residual conditions must also be met for this stability to maintain effective discrimination in the operating range.

2.3 Attention Models for Cross Condition Classification

Transformer encoders are ordered-token-processing modules that learn features via attention. Sequence Representation Studies [17] use positional information to maintain the relationship among local observations and their locations in a window. A token in eddy-current classification can be composed of adjacent spatial samples from several complex channels. Attention can be used to connect distant cues and distinguish between a compound pattern and a single general response. A small encoder is suitable for a moderate number of tokens in the observation window and is used to control changes in the input conditions. The capacity and training schedule of the raw and invariant versions can be kept the same.

Controlled Architecture Comparison [18] requires the representation path, objective and scoring process to be specified separately. Otherwise, the gain of normalisation will be mistakenly attributed to a larger encoder or an additional training view. Probability evaluation [19] is also different from a numerical confidence scale. Temperature scaling alters the distribution of probability mass among categories and keeps the largest logit for a positive temperature. Therefore, the current evaluation reports the matched Transformer variants, separate representation ablations, and calibration fitted on validation observations. Unseen lift-off observations are not used in the final classification, sensitivity and review analyses.

3 Cross Condition Learning with Lift Off Invariant Tokens

3.1 Eddy Current Encoding and Condition Partitioning

Each observation contains 96 ordered spatial samples over an 80 mm reference coordinate. Three complex channels represent nominal excitation settings of 25, 50 and 100 kHz, producing six real-valued input channels. The channel index identifies a response setting, while sample order retains location within the window. Real and imaginary components jointly preserve the modeled complex phase. These six trajectories enter training-derived channel standardization and subsequent window-level normalization before tokenization. The complex response relation is:

$$S_{f,t} = R_{f,t} + jI_{f,t} \tag{1}$$

The numerical response generator defines reference, compact, extended and compound source families with base width parameters 2.8, 3.8, 8.0 and 3.6 mm. A geometry-specific factor from 0.82 to 1.18 multiplies each width. A Gaussian derivative produces the real profile and an associated second-order profile supplies the quadrature component. The compound family adds a second lobe 11-18 mm away at relative strength 0.72; the reference family retains 0.07 of nominal response strength. Source amplitude ranges from 0.75 to 1.25 and center displacement from -9 to 9 mm. Lift-off then applies quadratic width broadening, channel-dependent exponential attenuation and phase rotation. Independent offsets, slow drift and noise complete the observations. Class labels therefore follow the source geometry before these condition transformations.

$$R = \frac{X - \mu}{\sigma} \tag{2}$$

Channel means and standard deviations are fitted only on training windows and retained for every later partition. The invariant pathway further centers each channel within its window and divides all six channels by one common root mean square energy, with stabilizer 0.01 inside the squared denominator. This reduces common offset and amplitude variation while retaining relative channel magnitude and complex phase structure. Four consecutive spatial positions across six real channels form one 24-value patch. Linear projection maps each patch into width 32, and a learned positional vector identifies its place among 24 ordered tokens.

$$h_t = Wr_t + p_t \tag{3}$$

The patch vector is r, token projection is W, and p denotes the learned positional embedding. Figure 1 shows the complex response channels, ordered window encoding and grouped condition-evaluation workflow. Two lift-off observations of each training geometry form paired views under independently drawn offsets and noise. Each held-out geometry supplies one seen-condition and one unseen-condition scoring view. Training-only channel statistics and validation-only checkpoint and temperature selection remain associated with their respective partition. The workflow links every representation and score to the source geometry and observation condition.

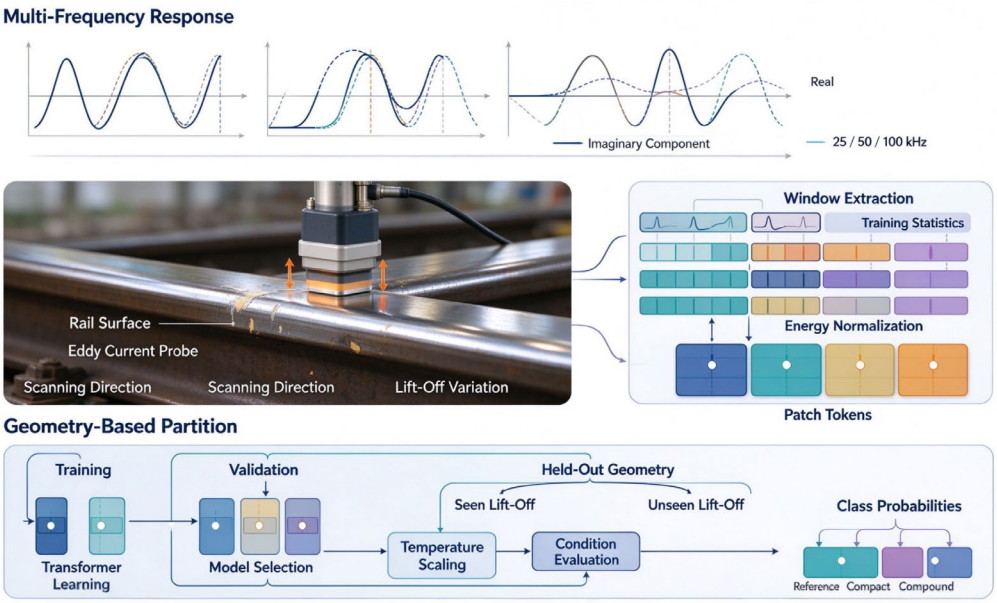


Figure 1. Eddy current observation encoding and grouped lift off evaluation workflow.

3.2 Lift Off Suppression and Transformer Aggregation

The encoder has two Transformer layers, four attention heads, model width 32 and feedforward width 64. Each layer uses a Gaussian error linear unit and dropout 0.05. Learned positional vectors retain spatial order across 24 tokens, and mean pooling yields one window representation. A linear head maps pooled features to four class logits. A separate projection head maps the 32 pooled features through a width-32 hidden layer to a 24-dimensional vector, then normalizes it to unit length. This projected vector z receives the lift-off consistency and class-conditioned contrastive objectives during training.

$$z = F(R) \tag{4}$$

The mapping F comprises token projection, Transformer encoding, pooling and the projection head. Two independently drawn lift-off observations from the training interval form a positive pair for each source geometry, with independently drawn offset and noise. Their unit projected vectors minimize mean cosine distance, promoting repeatability of geometry-linked features under the declared transformations. Figure 2 shows shared weights across the paired token and encoder branches. The projection branch carries the representation objectives, while pooled encoder features enter the four-class inference head. This separation identifies the representation target and the delivered class decision.

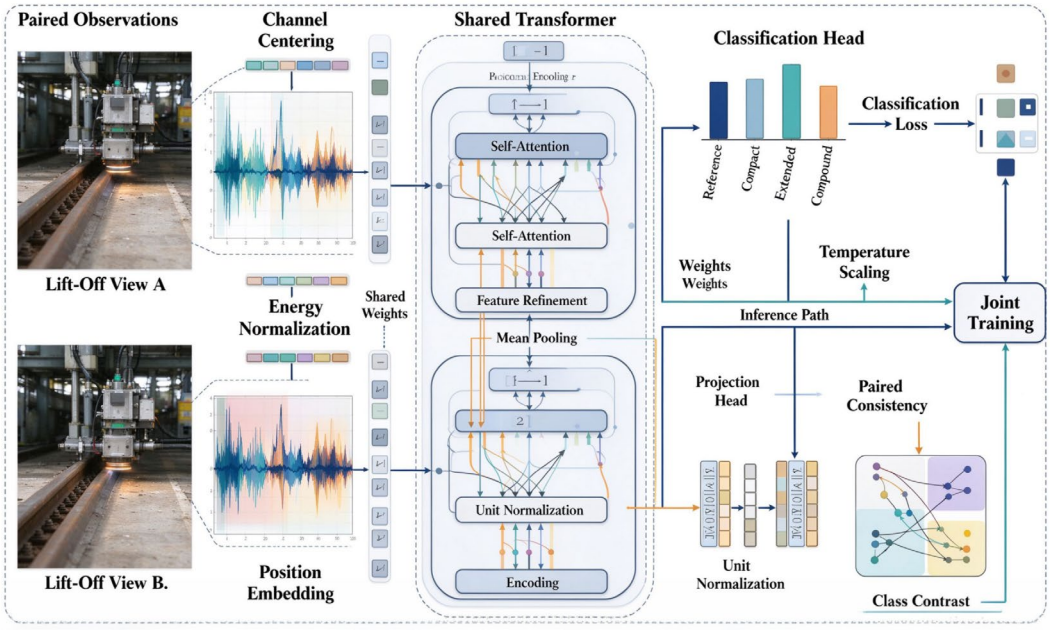


Figure 2. Shared Transformer encoding and paired lift off invariant representation learning.

$$L_p = 1 - z^T z' \tag{5}$$

The two paired branches share token projection, positional vectors, Transformer weights and classification head. Projected unit vectors receive paired and class-conditioned contrastive losses. Pooled encoder features supply class logits at inference. Cross-channel patches allow attention to associate spatial response morphology with the corresponding real-imaginary phase information. Each attention head forms query, key and value vectors from token states, and scales token similarity by the square root of its head width. The common encoder therefore supports both learned representation structure and the final four-class decision.

$$A = softmax(\frac{QK^{\mathrm{T}}}{\sqrt{d}})V \tag{6}$$

Attention weights aggregate token values according to spatial and channel-dependent similarity, including separated lobes of a compound response. The supervised contrastive objective uses projected vectors from both views of every source in a batch. Same-class observations form positive comparisons, self-comparisons are excluded, and all other observations remain in the normalized exponential denominator. The contrastive temperature is fixed at 0.15. Positive log probabilities are averaged to produce a batch term alongside source-specific paired consistency. Raw and normalized baselines train their pooled classification pathway; their auxiliary projection mapping remains at its initial state. The full objective trains that mapping directly, so projection statistics characterize different regularization pathways under the same encoder design.

3.3 Classification Objectives and Model Selection

Training averages classification cross entropy across both paired views and adds paired consistency and supervised contrastive losses with coefficients 0.4 and 0.1. Classification supervises pooled encoder features, while the auxiliary objectives act on projected unit vectors. AdamW uses learning rate 0.001, weight decay 0.0001, batches of 80 source geometries and gradient-norm clipping at two. Every compared encoder has a 32-epoch budget and a checkpoint selected by minimum validation classification loss. The geometry partition and optimization settings are common to the main comparison.

$$L = L_c + 0.4L_p + 0.1L_s \tag{7}$$

The symbols Lc, Lp and Ls denote classification, paired consistency and supervised contrastive losses. The pair-only variant sets the Ls coefficient to zero; the class-only variant sets the Lp coefficient to zero. Both retain the two-view classification loss, encoder and invariant preprocessing. The normalized Transformer instead uses a single energy-normalized view, while the raw Transformer uses training-channel standardization alone. Validation geometries are independent and drawn within the seen lift-off interval. This design separates the learned auxiliary objectives from architecture and preprocessing changes.

$$p = softmax(\frac{a}{T}) \tag{8}$$

Class logits a are divided by a positive temperature T before softmax. Validation negative log likelihood selects one of 101 temperatures spanning 0.5 to 3.0. Accuracy and macro-F1 use the largest class probability, while ten confidence bins, negative log likelihood and multiclass Brier score assess confidence. Retrospective review ranks held-out windows by increasing maximum probability and records the fraction of observed errors contained within each fixed budget. Temperature changes probability concentration while retaining the largest-logit class. Both the class output and its confidence therefore receive separate unseen-condition evaluation.

4 Generalisation to Unseen Lift Off Conditions

4.1 Signal Population and Condition Coverage

Geometry-level partitions separate fitting, validation and final scoring. Figure 3 presents the balanced class population and lift-off intervals. Figure 3(a) contains 300 training, 80 validation and 160 held-out geometries per response family, supporting class-specific and macro-averaged reporting. Figure 3(b) contrasts the seen interval of 0.5-2.5 mm with the unseen interval of 2.8-4.0 mm, leaving a 0.3 mm gap between them. Grouped generalization assessment [20] keeps all observations of one source geometry in the same partition. Seen and unseen held-out views therefore retain one source morphology and category while changing lift-off and independently drawn observation noise.

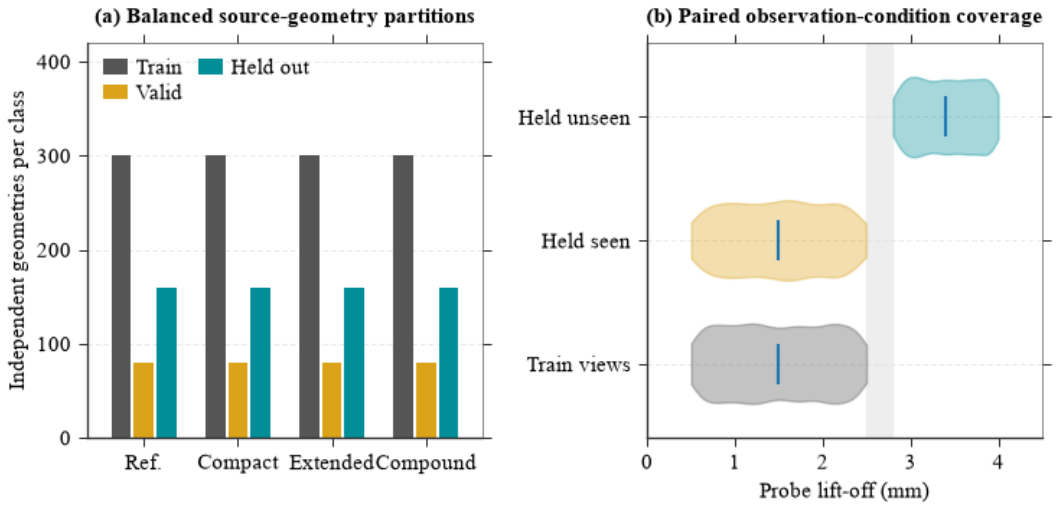


Figure 3. Class population and lift off coverage. (a) Geometry counts by category and partition. (b) Seen and unseen lift off intervals.

Figure 4 shows the changes in response generated by the observation model for the three nominal excitation channels. Figure 4(a) is the response amplitude at various lift-off values for the same geometry. The coefficients of the exponential attenuation are 0.27, 0.36 and 0.45 per millimetre, respectively; thus, the high-index channel has a higher loss in amplitude. Figure 4(b) is the phase rotation and the corresponding real-imaginary trajectories. The Phase Offset is 0.2 times the Channel Index plus 0.14 times the Lift-off (in mm). Multi-source representation analysis [21] can jointly observe these transformations, as a modification in one real-valued element may be caused by a rotation of the complex response. The relative channel factors of the comparison are also 1.00, 0.80 and 0.65. The above declared factors ensure the reproducibility of the size and phase relationship, and provide nuisance transformations for the embedding.

Figure 4(c) shows the change in the area of compact and extended indications with increasing lift-off. The initial widths of the two classes have different degrees of overlap and are thus distinguishable by morphology. A compound family can have a second lobe that is located independently at the point of division. Figure 4(d) shows the offsets, slow drift and additive noise at the same spatial location. The standard deviation of the main noise is 0.018 response units, and the drift slope is between -0.06 and 0.06 at a normalised coordinate. Controlled perturbation analysis [22] is related to the objective of observation change. Paired views are created from the same source geometry with different noise and offset. Therefore, the same embedding needs to aggregate the repeated lobe structure of observations with different sample values in the window. The Response Panel shows the numerical observation conditions of the classification experiment. The typical location of these is between -40mm and 40mm, and the spacing among samples is about 0.842mm. Uniformly sample all the channels and maintain the alignment of patch boundaries for local response comparison.

The data contain 1,200 training, 320 validation and 640 held-out source geometries. Geometry generation uses seeds 2951, 2952 and 2953 for those independent partitions, while training paired views use condition seeds 2954 and 2955. Data-partition method [23] treats the source identity as the grouping unit for repeated observations. Each held-out geometry produces one seen and one unseen scoring window, giving 640 windows per interval and 1,280 scored windows overall. All observations retain 96 spatial positions, three complex channels and 24 token patches. Geometry partitioning precedes view generation, preserving independence of source amplitude, center, width factor and compound-lobe separation across data roles.

The matched comparison includes a convolutional classifier, raw-input Transformer, energy-normalized Transformer and paired-objective variants. The convolutional classifier uses two 32-channel spatial convolutions with kernels seven and five, Gaussian error linear units and global average pooling. Transformer variants retain the same two-layer token encoder, class head and projection dimensions. Matched computational assessment [24] uses common labels, training-channel standardization and validation checkpoint criteria. Accuracy is the fraction correctly classified, macro-F1 averages four class scores, and class recall identifies response-family errors. The convolutional model stores 8,508 parameters, while each Transformer stores 20,636 including its projection head. Main results use seed 295; additional seeds assess fitted-objective variation.

Figure 5 compares classification in seen and unseen lift-off intervals. Figure 5(a) reports accuracy and macro-F1 for convolutional, raw, normalized and full-invariant models. In the unseen interval, the full model reaches 97.19% accuracy and macro-F1 0.9720, compared with 95.00% and 0.9500 for the raw Transformer. The normalized Transformer reaches 95.47% accuracy, isolating part of the preprocessing effect. Figure 5(b) displays the full-model confusion matrix over 160 held-out geometries per class. Reference, compact, extended and compound recalls are 100.00%, 90.62%, 98.75%

and 99.38%. Classification evaluation [25] accompanies the aggregate score with these class-specific outcomes: fifteen compact indications are assigned to compound, two extended to compound, and one compound to reference. The 18 errors concentrate in morphology-sensitive categories.

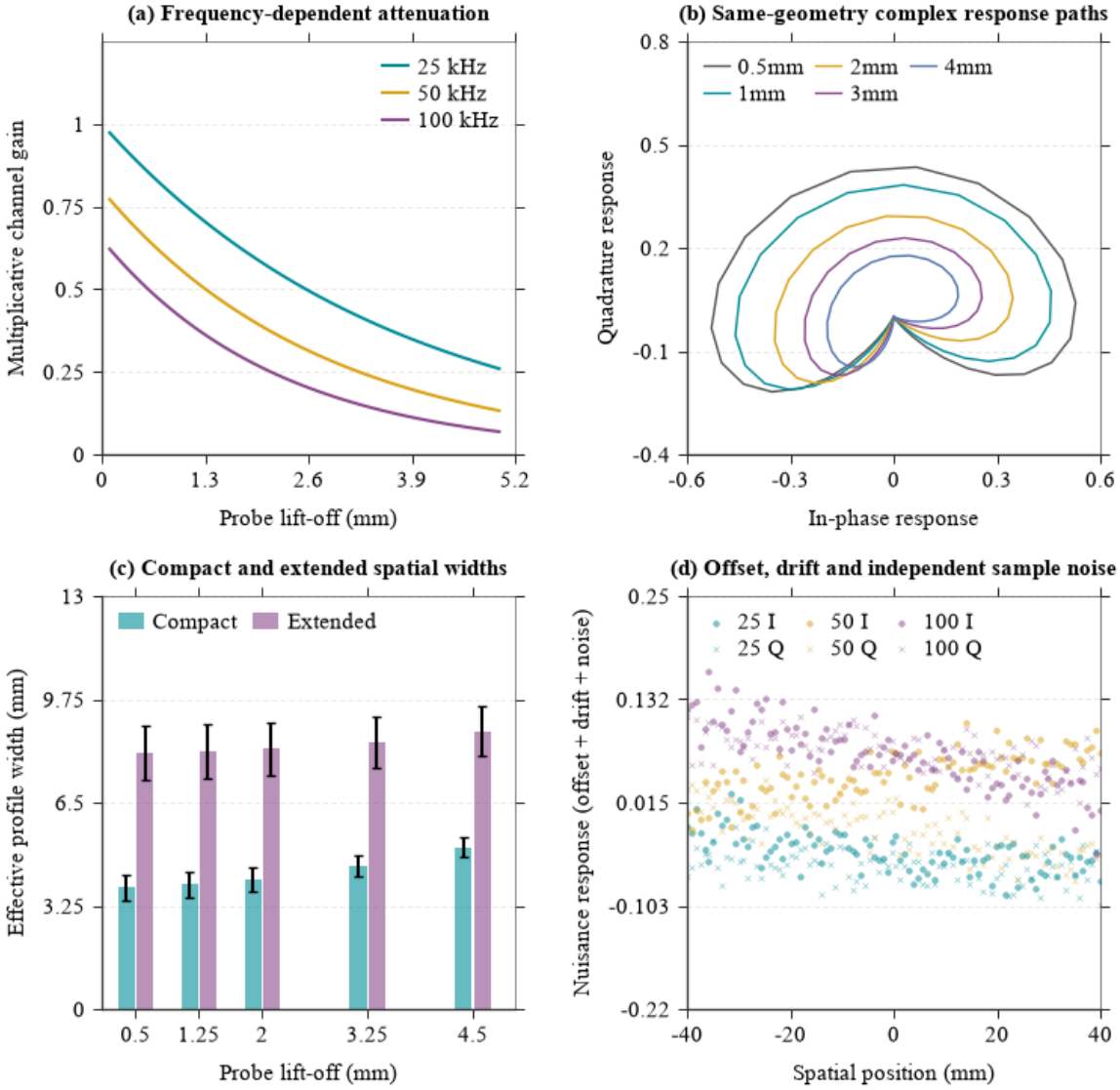


Figure 4. Multi-channel response changes across lift off conditions. (a) Magnitude attenuation. (b) Complex phase trajectories. (c) Spatial broadening. (d) Drift and additive noise.

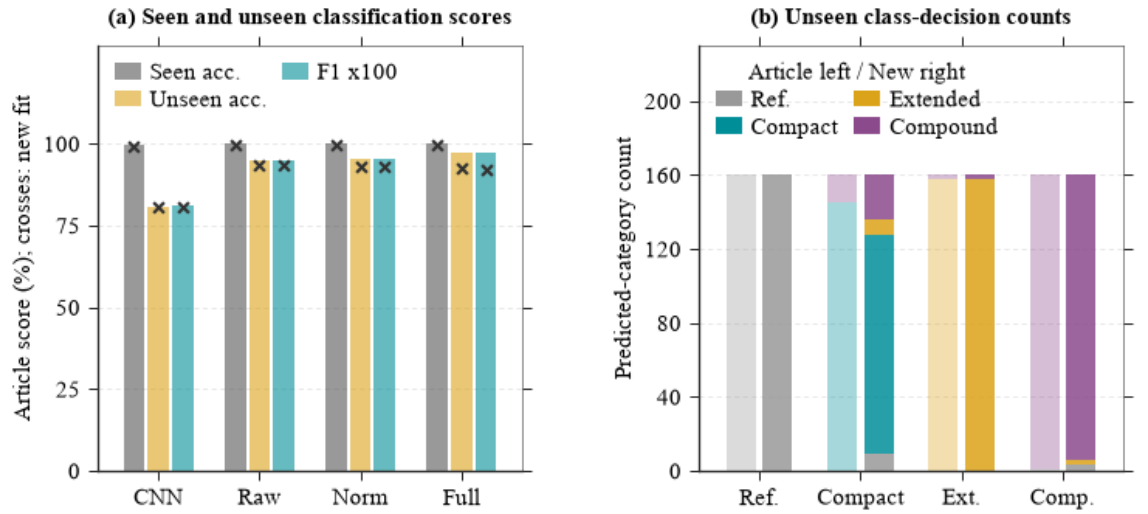


Figure 5. Classification across observation conditions. (a) Accuracy and macro F1 comparison. (b) Unseen lift off confusion matrix.

Table 1 presents the numerical cross-condition comparisons of seen accuracy, unseen accuracy and unseen macro-F1 for the five configurations. The additional pair-only configuration has the same normalisation and two-view class supervision as the full model, and can be used to evaluate the impact of class-conditioned contrastive organisation. Benchmark reporting [26] associates such a table with a specific condition interval and checkpoint rule. Thus, each row uses the same 640 seen and 640 unseen observations from the same held-out geometries. The first is called the ratio, and the macro-F1 is between 0 and 1. The difference between the two accuracy columns is the condition sensitivity of the fitted classifiers. The rows can be used to identify a model with good seen-condition performance that does not degrade after observation transformation. The table is added to show the general effect of each representation path in the scoring population on the class matrix.

Table 1. Classification comparison across seen and unseen lift off intervals.

Model	Seen accuracy (%)	Unseen accuracy (%)	Unseen macro F1
Convolutional classifier	99.69	80.78	0.8114
Raw Transformer	100.00	95.00	0.9500
Normalised Transformer	100.00	95.47	0.9543
Pair only objective	100.00	97.81	0.9782
Full invariant model	100.00	97.19	0.9720

Figure 6 summarizes projected-representation measurements on held-out source geometries. Figure 6(a) compares paired embeddings from lift-off intervals 0.6-1.0 mm and 3.2-3.8 mm. Mean cosine similarity is 0.9226 for raw Transformer, 0.9479 for normalized Transformer and 0.9655 for the full objective. Figure 6(b) reports a between-to-within class-center distance ratio of 4.7319 for raw embeddings and 14.6969 for full-invariant embeddings. Representation assessment [27] treats same-geometry similarity as repeatability and class-center separation as discrimination among source geometries. Both statistics use 24-dimensional unit projected vectors and complement the classification scores.

Figure 6(c) evaluates remaining lift-off information with a linear ridge probe fitted to projected training embeddings. Cross-validation chooses regularization from {0.01, 0.1, 1.0, 10.0}. Seen-interval held-out R squared is 0.3491 for raw Transformer and 0.0684 for the invariant model. Figure 6(d) compares repeatability and retained category discrimination under common observation pairs. Probing methodology [28] interprets the lower coefficient as reduced linearly accessible lift-off information within this specified projection and condition interval. Paired cosine similarity and class separation assess complementary representation properties. Probe fitting, validation and scoring remain within the seen range, while cross-condition classification is assessed separately.

The 640 unseen-condition windows contain 18 classification errors. A displaced compact response can resemble a compound response with a weak secondary lobe, and an extended response can overlap a broadened compact pattern. Error analysis in sequence classification [29] links these confusions to ambiguous spatial evidence under attenuation and broadening. The four source families occur equally often in this numerical test, so aggregate accuracy describes a balanced category population. Inspection settings with different reference prevalence require their own operating frequencies when interpreting class probabilities and review budgets. The saved confusion counts retain the family-specific error structure for such assessment.

4.3 Sensitivity Ablation and Probability Reliability

Figure 7 compares the representation objectives under a shared encoder and energy-normalized input. Figure 7(a) reports unseen accuracy of 97.81% for pair-only, 98.28% for class-only and 97.19% for the full objective. Pair-only aligns observations of one geometry, while class-only organizes different geometries by category. Figure 7(b) reports 95.47% for the normalized single-view Transformer. Component analysis [30] distinguishes these paths because two-view variants also average classification loss across their paired inputs. Under the stated primary seed and training budget, both single-term variants exceed the full objective on unseen accuracy. Their metric-specific ranking clarifies the trade-off between representation regularization and the delivered class decision.

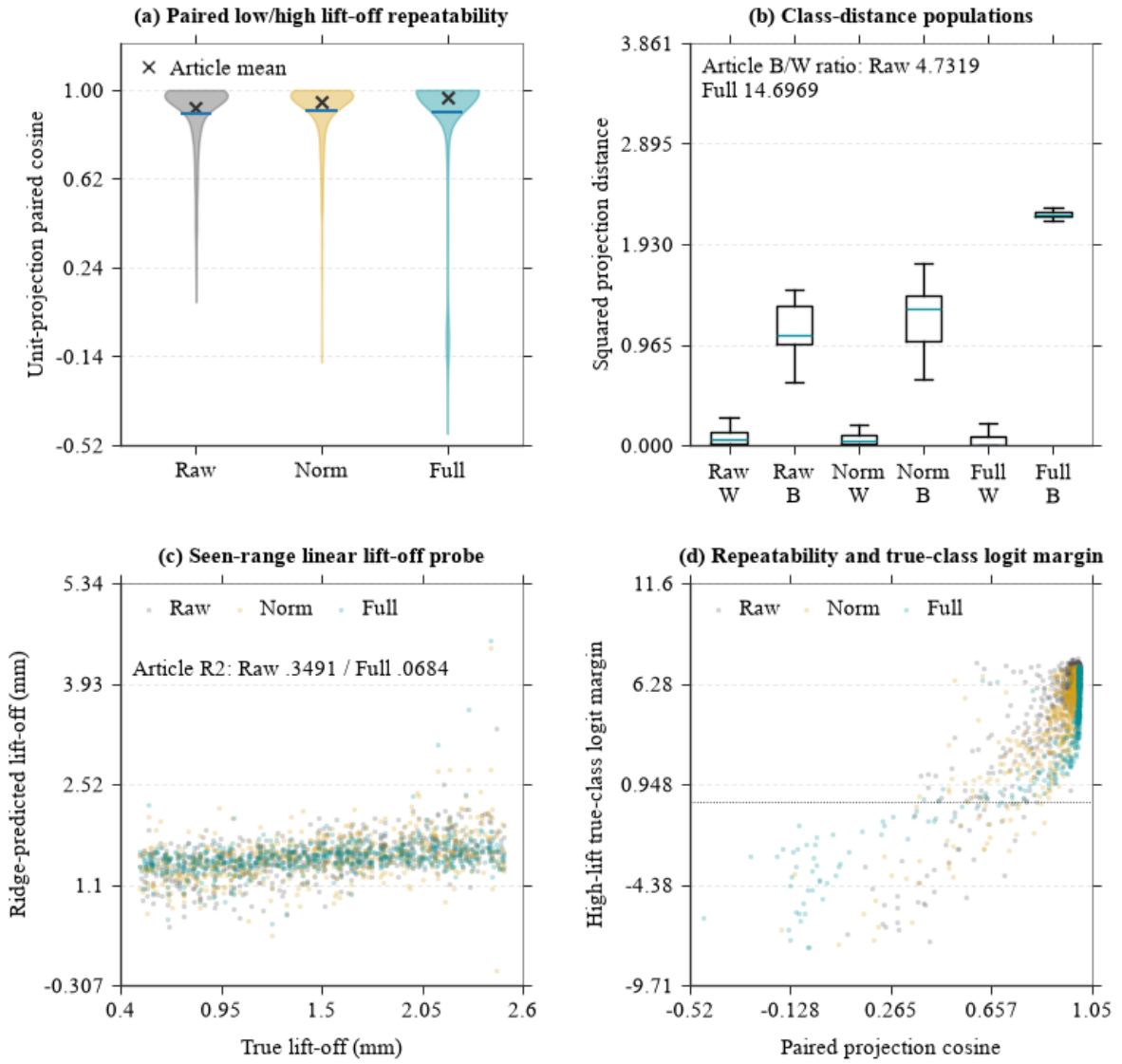


Figure 6. Representation repeatability and retained class information. (a) Paired embedding similarity. (b) Within and between class distances. (c) Linear lift off probe. (d) Repeatability and class separation comparison.

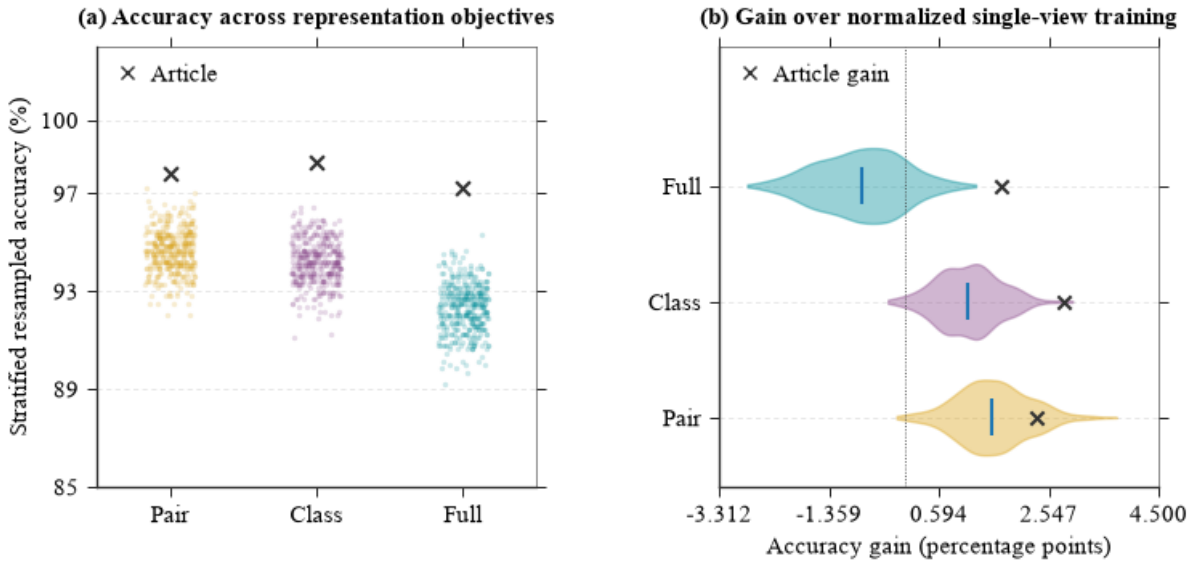


Figure 7. Contributions of representation objectives. (a) Pair and class conditioned objective comparison. (b) Comparison with the normalised single view Transformer.

Figure 8 presents disturbance, probability and review assessments for retained classifiers. Figure 8(a) reports full-model accuracy 100.00% at lift-off 0.1-0.4 mm and 73.28% at 4.1-5.0 mm; the normalized Transformer reaches 74.22% in the latter interval. Figure 8(b) applies either additive noise with standard deviation 0.040 or removal of ten central spatial samples across all six real channels. Robustness assessment [31] treats these as distinct changes to observation values and local evidence. Full-model accuracy is 76.56% under noise and 74.53% under the gap, while raw Transformer reaches 84.22% and 79.53%, respectively. Class labels and source geometry remain fixed within each intervention.

Figure 8(c) compares unseen-condition confidence before and after validation-selected temperature scaling. The chosen temperature is 0.5000, the lower edge of the search grid. Expected calibration error rises from 0.0155 to 0.0241, negative log likelihood from 0.0837 to 0.1304, and multiclass Brier score from 0.0423 to 0.0448. Calibration evaluation [32] therefore shows deterioration of unseen probabilities despite validation selection. Figure 8(d) ranks the same 640 windows by increasing calibrated maximum probability for retrospective review. Budgets of 32, 64 and 128 windows capture 50.00%, 88.89% and 94.44% of the 18 observed errors, equivalent to 9, 16 and 17 errors. Temperature scaling preserves class predictions but changes their probability quality and ordering-related confidence values.

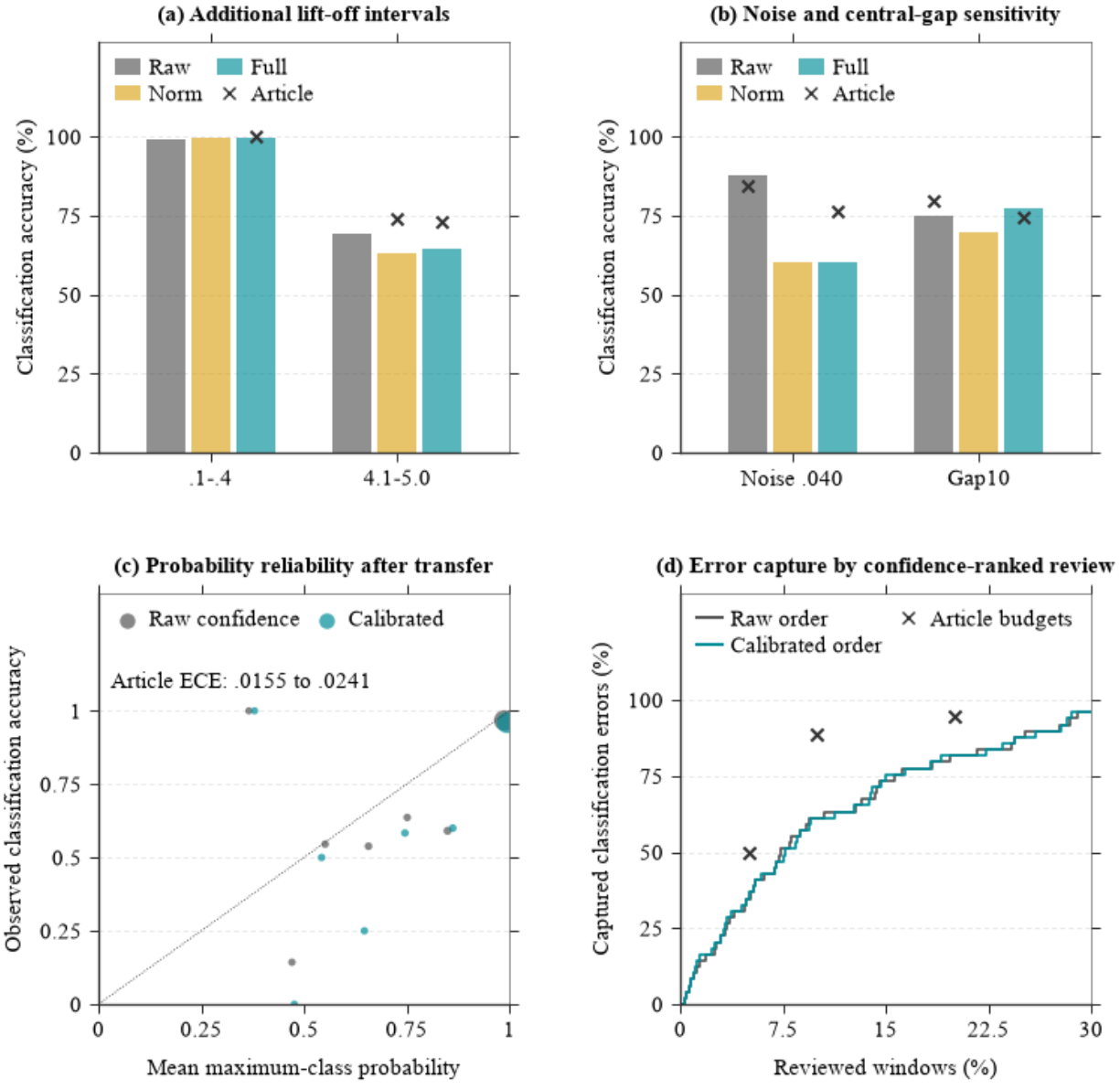


Figure 8. Sensitivity and probability-based review. (a) Additional lift off intervals. (b) Noise and spatial gap sensitivity. (c) Confidence reliability and calibration. (d) Error capture under review budgets.

Independent initialization seeds 295, 296 and 297 give unseen accuracies from 97.19% to 98.59%, with mean 97.76% under the same geometry population and 32-epoch schedule. Repeated-training assessment [33] describes this fitted-model variability. The temperature and review analyses use the primary seed 295 checkpoint. Selective prediction research [34] interprets review capture at explicit budgets: the 32-, 64- and 128-window groups contain 9, 16 and 17 of

the 18 unseen errors. A remaining error lies outside the 20% review budget. Stored category predictions, maximum probabilities and reference labels support reconstruction of the review curve under the balanced test population.

Lift-off tolerance is assessed through distinct outputs: classification accuracy measures category decisions, paired cosine similarity measures same-geometry repeatability, and ridge probing measures one form of linearly accessible condition information. Reproducible model evaluation [35] links these metrics to the declared attenuation, phase-rotation and broadening operators, geometry partitions and retained training settings. Additional high lift-off, noise and central-gap interventions characterize the same fixed encoder under changed evidence. The 20,636-parameter model processes 24 ordered tokens per window. Measured rail signals and their operating prevalence would provide external tests of category, representation and confidence transfer.

5 Conclusion

This study evaluated a Transformer classifier for four synthetic rail eddy-current response families under changing probe lift-off. Three complex channels were encoded as ordered spatial tokens, and a shared two-layer encoder integrated local morphology with window-level relationships. Training-derived standardization, window centering and common energy normalization controlled input scaling. Paired geometry consistency and supervised class contrast organized projected embeddings, while pooled encoder features supplied class predictions. Independent geometry partitions supported seen- and unseen-condition assessment.

First, on 640 unseen-condition geometries at lift-off 2.8-4.0 mm, the full model reaches 97.19% accuracy and macro-F1 0.9720, compared with 95.00% and 0.9500 for raw Transformer; class-only training reaches higher accuracy 98.28%. Second, full-model paired cosine similarity is 0.9655 versus raw 0.9226, and seen-range lift-off ridge-probe R squared falls from 0.3491 raw to 0.0684 invariant. Third, the retained full model reaches 73.28% accuracy at 4.1-5.0 mm, 76.56% under added noise and 74.53% under a central gap. These findings distinguish classification, representation structure and sensitivity to changed observations.

Validation-selected temperature 0.5 increases unseen ECE from 0.0155 to 0.0241 and NLL from 0.0837 to 0.1304; confidence transfer is therefore reported separately from class accuracy. Retrospective review of 32, 64 and 128 of 640 windows captures 9, 16 and 17 of 18 observed errors, respectively. Three full-objective initializations yield mean unseen accuracy 97.76% across the stated synthetic population. The conclusions apply to the declared response operators, balanced source families and grouped geometry split. Future measured-rail studies should test varied materials, probe wear, inspection speed, class prevalence and probability calibration under operational lift-off changes.

6. Declarations

6.1 Author Contributions

Mesut Rauf contributes to conceptualization, methodology, software, validation, analysis, investigation, data collection, draft preparation, manuscript editing, visualization, supervision, project administration, and funding acquisition. Oktay Savaş contributes to validation, analysis, investigation, data collection, draft preparation. All authors have read and agreed with the manuscript before its submission and publication.

6.2 Funding and Acknowledgments

This research received no specific financial support from any funding agency.

6.3 Institutional Review Board Statement

Not applicable.

6.4 Informed Consent Statement

Not applicable.

6.5 Date Availability

Not applicable.

6.6 Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

6.7 Declaration of generative AI and AI-assisted technologies

During the preparation of this manuscript, the authors utilized Grammarly for linguistic polishing, logical organization and draft revision. The generative AI tool was not involved in research design, data collection, data analysis, or the equation of core conclusions. All outputs generated by AI were reviewed, revised and verified manually by the authors. The authors take full responsibility for all content of this manuscript.

7. References

[1] Quercio, M., Santoro, L., Sesana, R., & Riganti Fulginei, F. (2025). Advanced feature analysis of eddy current testing signals for rail surface defect characterization. IEEE Access, 13, 141156–141169. https://doi.org/10.1109/ACCESS.2025.3597079

[2] Shao, Y., Xia, Z., Ding, Y., Crocker, B., Saunders, S., Bai, X., Peyton, A., Conniffe, D., & Yin, W. (2024). Eddy current sensor array for electromagnetic sensing and crack reconstruction with high lift-off in railway tracks. Sensors, 24(13), 4216. https://doi.org/10.3390/s24134216

[3] Ma, Q., Tian, G., Gao, B., Robinson, M., Yang, C., & Ibrahim, E. T. (2024). Novel common-differential inductance coils with dual signal conditionings for separation of lift-off and defects. IEEE Sensors Journal, 24(2), 2055–2065. https://doi.org/10.1109/JSEN.2023.3336066

[4] Rath, M., & Condurache, A. P. (2024). Boosting deep neural networks with geometrical prior knowledge: A survey. Artificial Intelligence Review, 57, 95. https://doi.org/10.1007/s10462-024-10722-5

[5] Khoee, A. G., Yu, Y., & Feldt, R. (2024). Domain generalization through meta-learning: A survey. Artificial Intelligence Review, 57, 285. https://doi.org/10.1007/s10462-024-10922-z

[6] Sasse, L., Nicolaisen-Sobesky, E., Dukart, J., Eickhoff, S. B., Götz, M., Hamdan, S., Komeyer, V., Kulkarni, A., Lahnakoski, J. M., Love, B. C., Raimondo, F., & Patil, K. R. (2025). Overview of leakage scenarios in supervised machine learning. Journal of Big Data, 12, 135. https://doi.org/10.1186/s40537-025-01193-8

[7] Hu, H., Wang, X., Zhang, Y., Chen, Q., & Guan, Q. (2024). A comprehensive survey on contrastive learning. Neurocomputing, 610, 128645. https://doi.org/10.1016/j.neucom.2024.128645

[8] Zhang, K., Wen, Q., Zhang, C., Cai, R., Jin, M., Liu, Y., Zhang, J. Y., Liang, Y., Pang, G., Song, D., & Pan, S. (2024). Self-supervised learning for time series analysis: Taxonomy, progress, and prospects. IEEE Transactions on Pattern Analysis and Machine Intelligence, 46(10), 6775–6794. https://doi.org/10.1109/TPAMI.2024.3387317

[9] Balanya, S. A., Maroñas, J., & Ramos, D. (2024). Adaptive temperature scaling for robust calibration of deep neural networks. Neural Computing and Applications, 36(14), 8073–8095. https://doi.org/10.1007/s00521-024-09505-4

[10] Li, Y., Xia, Z., She, S., Shao, Y., Yang, Y., & Yin, W. (2025). A Mobius transformation-based liftoff effect reduction method for crack classification and prediction in eddy current testing. NDT & E International, 149, 103261. https://doi.org/10.1016/j.ndteint.2024.103261

[11] Cho, Y. S., & Kim, S. B. (2024). Supervised contrastive learning for multisensor signals classification in automobile engine manufacturing. IEEE Transactions on Industrial Informatics, 20(5), 7764–7776. https://doi.org/10.1109/TII.2024.3363189

[12] Abdullahi, M., Alhussian, H., Aziz, N., Abdulkadir, S. J., Khairy, S. O. F., Bala, A., Adamu, S., & Yunus, R. B. (2026). A systematic literature review and taxonomy of normalization techniques in training DNNs for time-series data. Neurocomputing, 696, 134095. https://doi.org/10.1016/j.neucom.2026.134095

[13] Meng, B., Zhuang, Z., Ma, J., & Zhao, S. (2025). A pulsed eddy current data-driven interpretable neural network for metallic foreign objects' characteristics identification. IEEE Sensors Journal, 25(19), 35835–35843. https://doi.org/10.1109/JSEN.2025.3598371

[14] Uelwer, T., Robine, J., Wagner, S. S., Höftmann, M., Upschulte, E., Konietzny, S., Behrendt, M., & Harmeling, S. (2025). A survey on self-supervised methods for visual representation learning. Machine Learning, 114, 111. https://doi.org/10.1007/s10994-024-06708-7

[15] Cui, J., Zhong, Z., Tian, Z., Liu, S., Yu, B., & Jia, J. (2024). Generalized parametric contrastive learning. IEEE Transactions on Pattern Analysis and Machine Intelligence, 46(12), 7463–7474. https://doi.org/10.1109/TPAMI.2023.3278694

[16] Zhong, Y., Zhou, W., & Wang, Z. (2025). A survey of data augmentation in domain generalization. Neural Processing Letters, 57, 34. https://doi.org/10.1007/s11063-025-11747-9

[17] Abbasi, M., & Saeedi, P. (2024). Enhancing multivariate time series classifiers through self-attention and relative positioning infusion. IEEE Access, 12, 67273–67290. https://doi.org/10.1109/ACCESS.2024.3397783

[18] Lones, M. A. (2024). Avoiding common machine learning pitfalls. Patterns, 5(10), 101046. https://doi.org/10.1016/j.patter.2024.101046

[19] Wu, Y., Shi, B., Dong, B., & Zheng, Q. (2026). TransTS: An adaptive post-hoc method for probability calibration under label noise. Science China Information Sciences, 69, 122105. https://doi.org/10.1007/s11432-025-4576-9

[20] Apicella, A., Isgrò, F., & Prevete, R. (2025). Don't push the button! Exploring data leakage risks in machine learning and transfer learning. Artificial Intelligence Review, 58, 339. https://doi.org/10.1007/s10462-025-11326-3

[21] Migus, P. (2026). Newton–Puiseux analysis for interpretability and calibration of complex-valued neural networks. Neural Networks, 195, 108172. https://doi.org/10.1016/j.neunet.2025.108172

[22] Ilbert, R., Hoang, T. V., & Zhang, Z. (2024). Data augmentation for multivariate time series classification: An experimental study. In 2024 IEEE 40th International Conference on Data Engineering Workshops (ICDEW) (pp. 128–139). IEEE. https://doi.org/10.1109/ICDEW61823.2024.00023

[23] Babu, M. A. A., Pandey, S. K., Durisic, D., Koppisetty, A. C., & Staron, M. (2025). D-LeDe: A data leakage detection method for automotive perception systems. In Proceedings of the 11th International Conference on Vehicle Technology and Intelligent Transport Systems (pp. 210–221). SciTePress. https://doi.org/10.5220/0013476700003941

[24] Mohammadi Foumani, N., Miller, L., Tan, C. W., Webb, G. I., Forestier, G., & Salehi, M. (2024). Deep learning for time series classification and extrinsic regression: A current survey. ACM Computing Surveys, 56(9), Article 217, 1–45. https://doi.org/10.1145/3649448

[25] Opitz, J. (2024). A closer look at classification evaluation metrics and a critical reflection of common evaluation practice. Transactions of the Association for Computational Linguistics, 12, 820–836. https://doi.org/10.1162/tacl_a_00675

[26] Middlehurst, M., Schäfer, P., & Bagnall, A. (2024). Bake off redux: A review and experimental evaluation of recent time series classification algorithms. Data Mining and Knowledge Discovery, 38(4), 1958–2031. https://doi.org/10.1007/s10618-024-01022-1

[27] Klabunde, M., Schumacher, T., Strohmaier, M., & Lemmerich, F. (2025). Similarity of neural network models: A survey of functional and representational measures. ACM Computing Surveys, 57(9), Article 242, 1–52. https://doi.org/10.1145/3728458

[28] Waldis, A., Perlitz, Y., Choshen, L., Hou, Y., & Gurevych, I. (2024). Holmes ⌕ A benchmark to assess the linguistic competence of language models. Transactions of the Association for Computational Linguistics, 12, 1616–1647. https://doi.org/10.1162/tacl_a_00718

[29] Li, T., Guo, X., & Ji, C. (2025). A literature survey of shapelet quality measures for time series classification. PeerJ Computer Science, 11, e3115. https://doi.org/10.7717/peerj-cs.3115

[30] Liu, M., Sheng, H., Zhang, N., Zhao, P., Yi, Y., Jiang, Y., & Dai, J. (2024). DSDCLNet: Dual-stream encoder and dual-level contrastive learning network for supervised multivariate time series classification. Knowledge-Based Systems, 292, 111638. https://doi.org/10.1016/j.knosys.2024.111638

[31] Dong, C. G., Li, Z. D., Zheng, L. N., Chen, W., & Zhang, W. E. (2024). Boosting certificate robustness for time series classification with efficient self-ensemble. In Proceedings of the 33rd ACM International Conference on Information and Knowledge Management (pp. 477–486). ACM. https://doi.org/10.1145/3627673.3679748

[32] Dong, J., Jiang, Z., Pan, D., Chen, Z., Guan, Q., Zhang, H., Gui, G., & Gui, W. (2025). A survey on confidence calibration of deep learning-based classification models under class imbalance data. IEEE Transactions on Neural Networks and Learning Systems, 36(9), 15664–15684. https://doi.org/10.1109/TNNLS.2025.3565159

[33] Vos, G., van Eijk, L., Sarnyai, Z., & Rahimi Azghadi, M. (2025). Stabilizing machine learning for reproducible and explainable results: A novel validation approach to subject-specific insights. Computer Methods and Programs in Biomedicine, 269, 108899. https://doi.org/10.1016/j.cmpb.2025.108899

[34] He, Z., Li, Z., Dong, J., Liu, X., Jiang, M., & Hu, L. (2025). Conjunction subspaces test for conformal and selective classification. Information Sciences, 707, 122037. https://doi.org/10.1016/j.ins.2025.122037

[35] Semmelrock, H., Ross-Hellauer, T., Kopeinik, S., Theiler, D., Haberl, A., Thalmann, S., & Kowald, D. (2025). Reproducibility in machine-learning-based research: Overview, barriers, and drivers. AI Magazine, 46(2), e70002. https://doi.org/10.1002/aaai.70002