Implementation Experience of Automatic Pavement Raveling (Loss of Aggregate) Classification Using Machine Learning

Xinan Zhang[1], Bingjie Lu[1], Yi-Chang (James) Tsai[1], and Alex Middleton[2]
[1] Georgia Institute of Technology, Atlanta, GA | [2] Mississippi Department of Transportation (MDOT), Jackson, MS

RESEARCH BACKGROUND & PROBLEM STATEMENT

Pavement raveling (progressive loss of aggregate particles from the asphalt surface) is one of the most prevalent distresses affecting Open-Graded Friction Course (OGFC) pavements widely used across Florida, Georgia, Mississippi, and many other states. Raveling compromises skid resistance and pavement surface integrity, and if left untreated, accelerates deterioration and escalates rehabilitation costs.

Prior research has established a strong technical foundation. The National Cooperative Highway Research Program (NCHRP) IDEA Project 163 and a subsequent Florida Department of Transportation (FDOT) project successfully developed and validated machine learning (ML)-based raveling detection algorithms using three-dimensional (3D) Laser Crack Measurement System (LCMS) range imagery and Random Forest classification. More than 70% of state Departments of Transportation (DOTs) now routinely collect 3D pavement surface data, meaning most agencies already possess the foundational data infrastructure for automated detection.

Despite this demonstrated technical capability, implementing these models in operational DOT environments remains challenging. The barriers arise not from model accuracy alone, but from:

- Data readiness: lack of representative, well-balanced annotated datasets
- Annotation practices: inter-rater variability that undermines model reliability
- Quality assurance (QA): absence of structured QA procedures before operational use
- Organizational capacity: highly variable levels of artificial intelligence (AI) expertise across agencies

STUDY APPROACH: SYSTEM-LEVEL LIFE-CYCLE FRAMEWORK

This work presents implementation experiences of automatic raveling classification using ML, based on pilot applications conducted with FDOT and the Mississippi Department of Transportation (MDOT). Rather than proposing new algorithms, this work emphasizes a system-level, life-cycle approach to AI-enabled problem solving, covering the complete workflow from raw data to maintenance decisions:

Each stage of this framework addresses a distinct implementation challenge. Data selection establishes the foundation: because real network datasets are dominated by undistressed surfaces, agencies must deliberately curate a representative subset covering the full severity spectrum, ensuring that moderate and high-severity cases are adequately represented for model training.
Annotation and calibration determine model reliability. Before formal labeling begins, all annotators participate in a calibration session to align on severity definitions and resolve boundary cases through consensus. The annotation output is a comma-separated values (CSV) file mapping each image filename to a severity label (0 through 3), fully compatible with the RF training pipeline.
Model training or deployment depends on agency capacity. Agencies with in-house expertise retrain the RF classifier using their own annotated data, retaining full control over training configuration and model versioning. Agencies without AI staff deploy a pre-trained model through a structured application-level interface that guides the workflow without requiring Python expertise.
QA verifies predictions before operational use. Automated spatial and temporal consistency checks flag physically implausible severity patterns for targeted manual review. Confirmed model errors are corrected and fed back into the retraining loop. This QA step is mandatory before any results are used for maintenance planning.
Result interpretation translates image-level predictions into engineering-usable information. Predictions at 5-meter intervals are aggregated into segment-level condition zones through spatial clustering. Geographic Information System (GIS)-based visualization links severity predictions to milepost coordinates for network-level reporting and spatial analysis.
PMS integration connects outputs to maintenance decisions. Clustered severity zones are combined with independently assessed cracking and rutting conditions to determine treatment assignments. Outputs are adapted to agency-specific PMS deduct value formats and treatment trigger thresholds, enabling direct use in network-level maintenance programming without additional data conversion.

FINDING 1: DOT ADOPTION TIERS

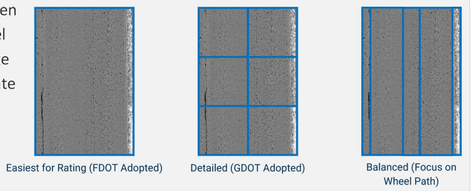
DOTs differ substantially in their ability to adopt ML-based raveling detection, and this work identifies two adoption tiers that should guide implementation strategy.
Tier A agencies have in-house AI capacity: agencies with internal data science or pavement engineering staff can implement and retrain the ML pipeline directly, retaining full control over training data, model configuration, and update schedules. This pathway is recommended for agencies that intend to maintain and iterate on the model over multiple survey cycles.
Tier B agencies are better served by application-level tools: agencies without dedicated AI staff achieve more reliable and sustainable outcomes by adopting pre-trained models through structured application-level tools with guided annotation workflows and built-in QA interfaces. This approach lowers the barrier to entry and reduces the risk of implementation failure due to technical capacity gaps.
Implementation strategy must be matched to agency capacity. A Tier B agency that attempts in-house retraining without adequate expertise is more likely to produce an unreliable model than one that adopts a pre-validated, structured deployment pathway.

FINDING 2: ANNOTATION QUALITY IS THE DOMINANT FACTOR

Annotation quality and consistency are the single most important factor determining implementation success, often outweighing the choice of ML model itself. Models trained on revised, higher-quality annotations of the same images consistently achieved higher accuracy than those trained on initial annotations. Three essentials for reliable annotation were established across the pilot studies:

- Rater calibration: all annotators must participate in a structured 30 to 60 minute calibration session before formal annotation begins, working through representative examples and reaching consensus on severity definitions for boundary cases (e.g., Level 1 vs. Level 2)
- Clear severity definitions: boundary cases must be explicitly illustrated with reference images agreed upon by the team, not left to individual interpretation
- Balanced datasets: no single severity class should account for more than approximately 60% of the total dataset; deliberate over-sampling of moderate and severe cases is typically required since undistressed surfaces structurally dominate any real network

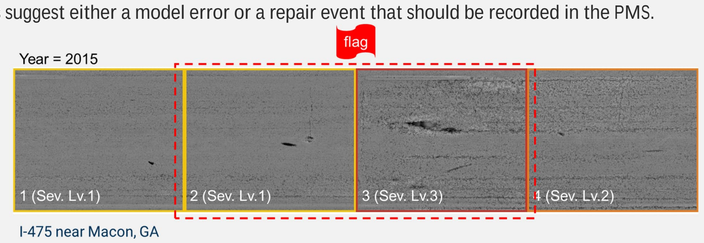
Three labeling granularity approaches have been used across the participating DOTs. Image-level labeling assigns one severity label per full range image. Sub-image-level labeling assigns separate labels to each of the six spatial tiles, capturing within-lane distress gradients. Column-level labeling assigns labels by wheel-path zone, balancing spatial detail with annotation efficiency.



Three labeling granularity approaches used across participating DOTs (image-level, sub-image-level, column-level)

FINDING 3: DOMAIN DIFFERENCES REQUIRE EXPLICIT QA

Cross-agency deployment without adaptation degrades model performance. Differences in sensing platforms, pavement textures, and mixed distress conditions across agencies introduce domain shift. Before results can be trusted for maintenance decision-making, explicit QA procedures are required and, in some cases, targeted retraining. Two automated consistency checks were implemented and validated across all pilot sites:
Spatial consistency: flags abrupt severity jumps of two or more levels between adjacent 5-meter images along a route. Such jumps are physically implausible under normal deterioration and typically indicate either a model error or a genuine physical boundary (e.g., a construction joint or patch). Flagged locations undergo targeted manual review.
Temporal consistency: flags severity changes of two or more levels at the same milepost across consecutive survey years. Under normal deterioration, pavement condition degrades gradually; large jumps suggest either a model error or a repair event that should be recorded in the PMS.



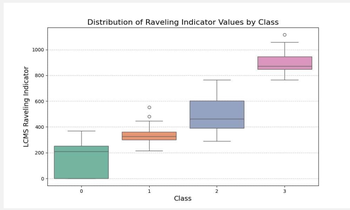
QA-triggered spatial discrepancy associated with a localized pavement feature (construction joint)

PILOT STUDY RESULTS

FDOT (Florida): Direct Deployment

FDOT direct deployment confirms cross-agency transferability when sensor conditions are consistent. An on-site visit to FDOT in December 2025 provided a 425-image LCMS dataset. The pretrained Random Forest classifier was applied directly without retraining. Results demonstrated strong positive correlation between RF model predictions and the FDOT operational LCMS raveling indicator, confirming that cross-agency transferability is achievable when both training and evaluation data are collected on the same LCMS platform.

ML outputs enabled targeted, cost-effective maintenance planning. Raveling-dominant sections were identified for targeted FC-5 surface-only replacement, avoiding the additional cost of deep milling where the underlying structure remained sound. Results were presented to FDOT senior management, contributing to executive-level adoption of AI-based pavement assessment.



Correlation between RF model predictions and LCMS raveling indicator

MDOT (Mississippi): Cross-Agency Domain Adaptation

A single day of targeted annotation was sufficient to overcome cross-agency domain shift at MDOT. The MDOT pilot tested cross-agency generalization using 1,287 images collected by LCMS 2.0 (different resolution from FDOT training data). Three sequential experiments quantified domain shift and the cost to close it:

- Experiment 1, direct deployment: positive but weaker correlation compared to FDOT. Domain shift from sensor specification differences, pavement texture, and climate reduced class separation at the Level 0/1 and Level 1/2 boundaries.
- Experiment 2, rapid annotation and retraining: 1,426 images annotated in a single day using stratified sampling. The domain-adapted RF model achieved reliable per-class accuracy on the held-out test set, as shown in Figure 26.
- Experiment 3, large-scale evaluation: the retrained model generalized consistently across the MDOT network. Persistent challenge: co-occurring distresses (raveling combined with cracking or patching) produced higher misclassification rates.

Confusion matrix of the retrained RF model on the MDOT held-out test set, demonstrating reliable per-class performance after domain adaptation

CONCLUSIONS & GUIDANCE FOR DOTS

This work provides a replicable pathway for DOTs seeking to adopt AI technologies for pavement raveling detection in a scalable and operationally meaningful manner.
Match implementation strategy to agency capacity. In-house teams with AI expertise can own the full ML pipeline and retrain directly across survey cycles, while agencies without AI staff achieve more reliable outcomes through pre-trained application-level tools with structured QA.
Invest in annotation quality above all else. Rater calibration sessions, clear severity definitions with reference images, and balanced datasets are non-negotiable prerequisites for model reliability. Models trained on revised, higher-quality annotations consistently outperform those trained on initial annotations of the same images, confirming that annotation quality outweighs model architecture as the determining success factor.
Build QA into standard operations. Spatial and temporal consistency checks catch domain-shift failures and model errors before they propagate into maintenance decisions. Flagged predictions enter a human review step: confirmed physical features are retained as section boundaries, while model errors are corrected and fed back into the retraining loop, progressively improving calibration over successive survey cycles.
Budget for a retraining cycle at new agencies. Cross-agency domain shift is real but surmountable. Differences in sensing platforms, pavement texture, and regional climate degrade direct deployment performance, but 1,426 images annotated in a single working day were sufficient to retrain the RF model to reliable per-class accuracy for full MDOT network deployment.