



BEYOND ACCURACY: EVALUATING EXPLAINABILITY AND ROBUSTNESS IN AI- BASED CIVIC ISSUE CLASSIFICATION

¹Harsha B K

¹Research Scholar/ B.Tech CSE

¹Department of Computer Science and Engineering,

¹Lovely Professional University, Phagwara, Punjab, India

Abstract : This paper develops a research framework for evaluating AI-based civic-issue classification beyond headline accuracy. Civic reporting systems can help public agencies identify potholes, waste accumulation, damaged streetlights, drainage failures, and other local infrastructure problems from citizen-submitted images. However, a model that performs well on a test set may fail when images are blurred, poorly illuminated, compressed, partially occluded, or captured in unfamiliar locations. A further concern is whether explanations offered by explainable artificial intelligence (XAI) methods faithfully represent the evidence used by a model. **Design/methodology/approach:** The paper proposes a reproducible comparative evaluation of image-classification models using a documented public dataset or a carefully governed civic-issue dataset. A conventional convolutional neural network is compared with a transfer-learning baseline and, where computational resources permit, a modern vision model. Evaluation includes macro-F1, per-class precision and recall, confusion matrices, calibration, inference latency, and controlled corruption tests. Explanation methods such as Grad-CAM are assessed using deletion/insertion or perturbation-based faithfulness checks and stability analysis rather than visual inspection alone. **Expected contribution:** The framework treats predictive performance, robustness, explanation quality, and operational efficiency as related but distinct dimensions. It proposes a reporting protocol that makes model limitations visible to researchers and public-sector users. No new experimental findings are claimed in this manuscript; numerical results must be produced after implementation and validation. The paper offers a feasible route from a student software project to a defensible empirical study, with relevance to responsible AI, computer vision, and digital public services.

IndexTerms - Explainable Artificial Intelligence, Computer Vision, Civic Issue Classification, Model Robustness, Grad-CAM, Responsible AI, Public Services, Image Classification

I. INTRODUCTION

Digital public-service systems increasingly rely on citizens to report local problems through websites and mobile applications. A report may include a short description, a location, a timestamp, and one or more photographs. Such systems can improve the visibility of problems that otherwise remain scattered across telephone calls, paper registers, and informal communication channels. Yet the volume and variety of incoming reports can make manual sorting slow and inconsistent. Image classification offers one way to assist triage by assigning an incoming report to a likely category, such as road damage, waste, drainage, or streetlight failure.

The motivation for this research is practical as well as technical. The Parivartan civic-issue reporting project provides a relevant application context: an automated system can analyse an uploaded image and help route a complaint to the appropriate department. In such a setting, an incorrect classification is not merely a benchmark error. It may delay a repair, direct a complaint to the wrong team, or cause a public employee to overlook an urgent issue. A system therefore needs to be evaluated for more than average accuracy on a static dataset.

Modern computer-vision systems can achieve strong results on benchmark data, but performance may degrade when deployment conditions differ from training conditions. Images submitted by the public vary in camera quality, distance, angle, lighting, weather, compression, and background clutter. A photograph of a pothole may contain vehicles, shadows, water, road markings, or construction equipment. A waste-related image may include several objects, and the most visually prominent object may not be the one relevant to the report. Distribution shift can therefore create a gap between laboratory performance and real-world reliability.

A second challenge concerns explainability. Public agencies and users may reasonably ask why a model assigned a category to a particular image. Visual explanation techniques, including class activation maps, can highlight image regions associated with a prediction. However, a heatmap that appears intuitive does not automatically establish that the explanation is faithful to the model's decision process. Explanations may be unstable under small input changes, may focus on spurious background features, or may offer a persuasive visual narrative without reliably describing the model's reasoning. Explanation quality must consequently be examined as an empirical question.

This paper proposes a structured evaluation of classification performance, robustness, explanation faithfulness, and operational cost. It does not assume that a single model will be best across all dimensions. A larger model may deliver better

clean-image performance but require more computing resources; a smaller model may be easier to deploy but more sensitive to certain corruptions. Likewise, a model with attractive heatmaps may fail perturbation-based explanation tests. The aim is to make such trade-offs explicit rather than collapse them into one score.

The research is guided by four questions. RQ1 asks how candidate models compare on overall and class-specific civic-issue classification. RQ2 asks how performance changes under controlled image corruptions. RQ3 asks whether explanations identify decision-relevant regions and remain stable under semantically minor changes. RQ4 asks how predictive performance and explanation quality trade off against inference latency and computational demands. These questions are intentionally narrow enough for a student-led study while remaining relevant to applied machine learning.

The proposed contribution has three parts. First, it defines an evaluation protocol suitable for civic image classification. Second, it combines corruption-based robustness tests with quantitative explanation checks. Third, it outlines reporting and governance practices that can make an AI-assisted complaint-routing tool more transparent. The remainder of the paper reviews related research, establishes the conceptual framework, specifies hypotheses and methodology, and discusses how results should be interpreted once experiments have been run.

II. BACKGROUND AND CONTEXT

Civic issue reporting is a classification problem embedded in a larger administrative workflow. The model's output is typically a proposed category rather than a final decision. A practical system may also need to detect duplicate reports, extract a location, estimate urgency, protect personal information, and provide a route for human review. This study isolates image classification because it is measurable and can be investigated with public datasets, while recognising that classification quality alone does not establish the effectiveness of the full service.

A useful taxonomy should be operationally meaningful. Categories might include potholes or road-surface damage, accumulated waste, blocked drainage, damaged public infrastructure, and other or uncertain. Categories must be defined before annotation, with guidance for ambiguous cases and multi-label images. If a photograph depicts both a pothole and standing water, the dataset protocol should specify whether annotators select the primary visible issue, assign multiple labels, or mark the case as ambiguous. Without a stable definition, measured model performance can reflect inconsistent annotation rather than algorithmic capability.

The system should be designed as decision support. A model can recommend a department and display a confidence estimate, but low-confidence or out-of-distribution examples should be routed to human review. This design is particularly important where the consequences of a mistake differ across classes. Misclassifying a minor litter report may be inconvenient, while misrouting a dangerous road defect could have greater consequences. The study therefore recommends reporting per-class errors and establishing a review policy rather than relying on a single aggregate score.

The context also introduces privacy and fairness considerations. Citizen photographs may inadvertently include faces, vehicle registration plates, private property, or location information. Data collection should minimise unnecessary personal information and follow the dataset's licence and applicable institutional requirements. If images are collected directly from people, informed consent, retention rules, secure storage, and a process for removal should be established. Public availability does not remove the need to assess whether a dataset is appropriate for the intended use.

Finally, civic reporting environments are geographically and temporally variable. Road materials, waste collection practices, signage, lighting, weather, and camera usage can differ across locations. A model trained on one city or one dataset may learn patterns that do not transfer. Robustness tests on synthetic corruptions are useful for controlled comparison, but they cannot replace evaluation on genuinely new locations. The study therefore treats corruption testing as one part of an evidence package, not proof of deployment readiness.

III. STATEMENT OF THE PROBLEM

Many image-classification studies report overall accuracy or F1 scores on held-out test sets, but these measures may conceal operational weaknesses. Aggregate scores can be dominated by common classes, leaving rare but important categories poorly served. A model may also exploit shortcuts such as background texture, image watermarks, or camera-specific patterns. When such correlations change in deployment, predictions can become unreliable even though the original test score was high.

In civic reporting, a classification error can influence workflow routing and response time. If the system cannot identify when it is uncertain, staff may trust a confident but incorrect recommendation. If explanations are provided without validation, users may interpret them as evidence of reliability even when the highlighted regions are irrelevant to the decision. Thus, there is a need to evaluate predictive quality, robustness, and explanation quality together while documenting computational cost.

The specific problem addressed is the lack of a compact, reproducible evaluation protocol that connects these dimensions for civic-issue image classification. The proposed study compares selected models under a common data split, measures performance on clean and corrupted images, evaluates explanations with quantitative tests, and records inference latency. The study does not claim that existing research entirely ignores these dimensions; rather, it aims to integrate them into a practical student-scale workflow with transparent reporting.

IV. RESEARCH GAP

First, model comparison and deployment robustness are often reported separately. A paper may establish a strong benchmark score without systematically showing how the model behaves under low illumination, blur, noise, compression, or partial occlusion. For a citizen-reporting application, these conditions are foreseeable rather than exceptional. An integrated experiment can reveal whether model rankings remain stable across conditions.

Second, explainability is sometimes evaluated primarily through qualitative figures. Heatmaps can be useful for communication, but visual plausibility is not a sufficient criterion for faithfulness. Quantitative perturbation tests and stability

analysis can provide complementary evidence about whether explanations track decision-relevant pixels and remain consistent when the input is minimally altered.

Third, practical systems face latency and resource constraints. A model's best accuracy may not justify a large inference cost if reports must be processed quickly on modest infrastructure. Reporting latency alongside predictive and explanation metrics can help identify viable trade-offs. Fourth, civic-issue taxonomies and datasets can be uneven, geographically specific, or weakly documented. A transparent protocol that records class distributions, data provenance, preprocessing, and limitations improves the interpretability of any results.

The intended gap is therefore methodological and applied: to demonstrate how a compact, reproducible evaluation can examine clean-image performance, corruption robustness, explanation faithfulness, explanation stability, and inference efficiency in one study. The paper's contribution will depend on the actual experiments and the extent to which the selected dataset supports these analyses.

V. OBJECTIVE OF THE STUDY

The primary objective is to develop and apply a reproducible framework for evaluating AI-based civic-issue image classification beyond conventional accuracy. The specific objectives are: (1) compare a small set of appropriate computer-vision models on a clearly defined civic-issue classification task; (2) measure class-specific performance and error patterns; (3) evaluate robustness under controlled image corruptions; (4) assess whether visual explanations are faithful and stable; (5) compare inference latency and model size; and (6) translate the findings into practical recommendations for human-in-the-loop public-service systems.

The study is not designed to claim that an algorithm can independently resolve civic complaints or determine the priority of public works. It evaluates a component that could assist routing. Any broader claim about service efficiency would require a field study involving users, public officials, and operational outcomes such as routing accuracy, time to assignment, duplicate handling, and complaint resolution.

VI. RESEARCH QUESTIONS AND HYPOTHESES

RQ1: How do the selected models compare on overall and per-class classification performance? RQ2: How much does performance degrade under controlled corruptions, and do model rankings differ by corruption type? RQ3: Do explanation methods highlight decision-relevant image regions, as assessed by perturbation-based faithfulness and stability measures? RQ4: What trade-offs exist among classification performance, robustness, explanation quality, and inference latency?

H1: Models will differ significantly in macro-F1 and class-specific recall on the clean test set. H2: All candidate models will experience measurable performance degradation under at least some corruption conditions, with the magnitude varying by corruption type and severity. H3: Qualitative plausibility of heatmaps will not always correspond to stronger quantitative faithfulness or stability scores. H4: Higher-capacity models may offer performance advantages on some conditions but will generally require greater computational resources than lighter baselines. These are hypotheses to test, not findings already established by this manuscript.

The hypotheses should be refined after a pilot run and a power or uncertainty analysis. If the dataset is small, the paper should report confidence intervals and avoid overstating statistical significance. Where the same test images are evaluated across models and corruption levels, comparisons should account for paired observations. The exact statistical tests should be selected before inspecting the final results to reduce post-hoc flexibility.

VII. THEORETICAL AND CONCEPTUAL FOUNDATION

The study draws on three complementary ideas: statistical generalisation, explainability as a property requiring evaluation, and human-centred decision support. Statistical learning theory motivates separating training, validation, and test data and examining whether performance persists beyond the observed sample. Robustness research highlights that small or structured changes in inputs can reveal weaknesses hidden by standard benchmark testing. Explainable AI research distinguishes the production of an explanation from evidence that the explanation is faithful to a model. Human-centred AI adds the principle that a technically accurate prediction is not automatically suitable for an operational workflow.

The conceptual model places model architecture and training procedure as explanatory factors, with clean-image performance, corruption robustness, explanation quality, and inference cost as outcome dimensions. Image corruption type and severity are experimental conditions. Class frequency, image source, and background context are potential confounders that must be documented. Explanation method is an additional factor because different methods can produce different maps for the same model and image.

The framework does not assume that explanation quality mediates classification accuracy. An explanation is an analysis of model behaviour, not necessarily a mechanism that improves prediction. Instead, explanation faithfulness and stability are evaluated as separate dimensions that may influence whether a model is appropriate for human review. Similarly, robustness is not treated as equivalent to fairness: corruption tests measure sensitivity to selected image transformations, while fairness requires separate analysis of groups and deployment contexts.

The proposed framework can be represented as a multi-criteria evaluation. The model is tested on clean images and transformed variants; prediction metrics quantify task performance; explanation metrics quantify selected aspects of model transparency; and latency or model size quantifies operational cost. The final comparison should use a scorecard rather than an opaque single composite score. If a composite is used, its weighting must be justified and sensitivity-tested.

VIII. LITERATURE REVIEW

Deep learning has substantially changed image classification. Convolutional neural networks learn hierarchical visual representations and have become standard baselines for supervised vision tasks. Transfer learning allows a model pretrained on a

large image corpus to be adapted to a smaller domain dataset, which can be valuable when labelled civic images are limited. However, transfer learning does not guarantee reliable domain adaptation. Pretrained models may rely on features that are useful in the source domain but less relevant to civic images, and performance depends on data quality, class definitions, and fine-tuning decisions.

The ImageNet Large Scale Visual Recognition Challenge helped establish large-scale benchmark evaluation as a central practice in computer vision (Russakovsky et al., 2015). Benchmark success is informative, but it is not equivalent to robustness under every deployment condition. Geirhos et al. (2020) discuss shortcut learning, where a model uses a decision rule that performs well on the observed data but fails under changed conditions. In a civic setting, a classifier might associate a category with background context or a particular image source instead of the issue itself. This motivates error analysis and tests that probe whether the model generalises beyond superficial cues.

Robustness studies investigate how models respond to distribution shift and corrupted inputs. Hendrycks and Dietterich (2019) introduced benchmark corruption types for evaluating common image-classification robustness, including noise, blur, weather effects, and digital distortions. Such controlled transformations allow repeatable comparisons and help identify sensitivity patterns. Nevertheless, synthetic corruptions have limits: a digitally blurred image is not identical to an image captured at night by a low-quality camera, and the set of transformations may not cover the full distribution of real civic submissions. Corruption tests should therefore be paired with qualitative inspection and, when feasible, external validation data.

Adversarial robustness is related but distinct. Adversarial examples are often constructed to induce a model error through carefully designed perturbations. The current study focuses primarily on common, non-adversarial image degradation because it is directly relevant to everyday reporting. It does not claim to provide a comprehensive security assessment. If the system later becomes a public endpoint, adversarial manipulation, data poisoning, abuse, and model extraction would require separate threat modelling and testing.

Explainable AI methods aim to make model behaviour more interpretable. Saliency methods estimate which input features influence a prediction, while class activation mapping methods highlight spatial regions associated with a class. Grad-CAM, introduced by Selvaraju et al. (2017), uses gradients flowing into a convolutional layer to produce a coarse localisation map for a target class. These maps can support debugging and communication, but their resolution and interpretation depend on architecture and implementation. A map may highlight a region correlated with the label without establishing causal importance.

A central issue in XAI is evaluation. Adebayo et al. (2018) proposed sanity checks for saliency maps, showing why explanations should be tested against model and label randomisation rather than accepted on appearance alone. Deletion and insertion tests examine how a model's output changes as highly ranked or low-ranked features are removed or added. Such tests have their own limitations, because image perturbation can create out-of-distribution inputs and the ranking can depend on the baseline or masking strategy. No single metric fully captures human interpretability, causal faithfulness, and usefulness in a workflow.

Ribeiro, Singh, and Guestrin (2016) introduced LIME as a model-agnostic approach that explains individual predictions through locally interpretable approximations. Lundberg and Lee (2017) introduced SHAP as a unified framework for feature attribution. These methods are influential but not interchangeable. LIME depends on the local perturbation distribution and surrogate model; SHAP explanations depend on the feature representation and assumptions used to estimate contributions. For image data, segmentation and feature grouping choices can materially affect the explanation. The proposed study therefore selects a method compatible with the candidate architectures and documents its assumptions.

Model calibration is another dimension of reliability. A classifier can be accurate yet overconfident, especially on inputs that differ from training examples. Guo et al. (2017) examine neural-network calibration and temperature scaling, illustrating that confidence quality is distinct from top-1 accuracy. For civic routing, confidence estimates can support an abstention policy, but confidence should not be treated as a guarantee that a prediction is correct. Calibration should be assessed on held-out data and, where possible, under distribution shift.

The human factors literature warns against assuming that adding an explanation automatically improves decisions. Users may over-trust a model when a visual explanation appears authoritative, or ignore useful predictions when explanations are inconsistent. The interface should make clear that a prediction is a recommendation, display uncertainty appropriately, and provide an easy route for correction. User studies are needed to establish whether explanations improve appropriate reliance, reduce handling time, or increase detection of errors. The present study evaluates technical explanation properties and identifies human evaluation as a subsequent stage.

Civic technology also intersects with responsible AI and public-sector accountability. Public-facing systems should be designed with transparency, privacy, accessibility, and human oversight. A classification model should not silently determine eligibility, impose penalties, or deprioritise residents without accountable review. Dataset documentation and model cards can make intended use, limitations, evaluation conditions, and risks more visible. Mitchell et al. (2019) propose model cards as a structured way to report model performance across relevant conditions, while Gebre et al. (2021) advocate datasheets that document dataset motivation, composition, collection, and recommended uses.

The reviewed literature points to a practical synthesis. Standard classification metrics are necessary but incomplete; corruption benchmarks reveal some aspects of robustness; explanation methods require sanity checks and quantitative evaluation; calibration can help manage uncertainty; and documentation is essential for responsible use. A student-scale civic classification study can contribute by bringing these elements into one protocol, carefully limiting claims to the conditions tested, and publishing enough detail for replication.

IX. RESEARCH METHODOLOGY

Research design: The study uses a quantitative experimental design based on supervised image classification. Candidate models are trained or fine-tuned under documented conditions and compared on identical test examples. Clean test performance is reported separately from performance under controlled corruption. Explanations are generated for a predefined subset of correct and incorrect predictions, and latency is measured using a consistent hardware and software environment. The experimental plan should be preregistered in a repository or dated protocol before the final test set is evaluated.

Dataset selection: The first choice should be a publicly available dataset whose licence permits research use and whose categories map reasonably well to civic issues. Road damage datasets such as RDD2020 may support the road-damage portion of the task, but they do not by themselves cover all civic categories such as waste and drainage. If no single dataset adequately covers the taxonomy, the study should narrow its claims or combine compatible datasets with explicit source labels and harmonised annotation rules. It must not describe a road-damage-only dataset as a comprehensive civic-issue dataset.

Data governance: Record the dataset name, version, access date, licence, source, class distribution, image resolution, annotation process, and any known collection biases. Remove duplicates and near-duplicates before splitting. If multiple images come from the same location, event, or video sequence, keep related images in the same split where possible to prevent leakage. Use training, validation, and test partitions, for example 70/15/15 where dataset size permits, but choose the split based on dataset structure rather than applying a fixed ratio mechanically. The test set must remain untouched during model selection.

Annotation and label quality: Where new labels are required, prepare a short annotation guide with class definitions and examples of ambiguous cases. Ideally, two annotators independently label a subset and disagreements are resolved through documented adjudication. Inter-annotator agreement can be reported where appropriate. If only existing labels are used, inspect label quality and document any mismatch between original labels and the research taxonomy. A category named “other/uncertain” should not become a dumping ground without a clear rule, because it can obscure errors and reduce interpretability.

Candidate models: A feasible comparison includes a compact CNN baseline, a transfer-learning model such as ResNet-50 or MobileNetV3, and an additional architecture only if resources allow. Model selection should be based on the task and compute budget, not on the expectation that the newest model must win. A pretrained model can be fine-tuned with a fixed training protocol, while the baseline provides a reference point. All architectures should use comparable data splits and evaluation conditions. If model-specific preprocessing differs, that difference should be documented and implemented correctly.

Training procedure: Specify image resizing, normalisation, augmentation, batch size, optimiser, learning rate, number of epochs, early-stopping criterion, random seed, and software versions. Augmentation should reflect plausible variation without changing the label semantics. For example, mild brightness variation may be appropriate, while a transformation that removes the road defect or changes the meaning of an image may not be. Hyperparameters should be selected using the validation set. Training should be repeated across multiple seeds if computationally feasible, with mean and variability reported rather than only the best run.

Clean-data metrics: Report overall accuracy but do not use it as the only measure. Macro-precision, macro-recall, and macro-F1 give each class equal weight and are useful when classes are imbalanced. Include per-class precision and recall, a confusion matrix, and support counts. If the task is multi-label, use metrics suitable for multi-label prediction and define the decision threshold. Report confidence intervals where feasible, using a method appropriate to the sampling design. Any class with very few examples should be clearly flagged as providing weak evidence.

Corruption protocol: Evaluate a predefined set of transformations such as Gaussian noise, motion blur, defocus blur, reduced brightness, contrast changes, JPEG compression, and partial occlusion. Select severity levels before testing and apply the same transformations to each model's test images. Preserve the clean test set as a separate condition. Report metric degradation relative to clean performance and examine whether specific classes are more vulnerable. Avoid choosing only transformations that favour one model. Synthetic corruption should be presented as a controlled stress test, not a direct estimate of field performance.

Explanation methods: Use Grad-CAM for compatible convolutional architectures, or select another method with a clear implementation and documented assumptions. For each selected example, generate the explanation for the predicted class and, where useful, the true class. Include correct, incorrect, high-confidence, and low-confidence examples. Quantitative evaluation can include deletion/insertion curves or other perturbation-based measures. If a segmentation mask of the issue is available for a subset, localisation overlap can be measured, but such overlap should not be treated as the only definition of explanation quality.

Explanation faithfulness and stability: A deletion test progressively removes features ranked as important and measures the change in the target score; an insertion test adds features according to the ranking. The interpretation of scores depends on the masking baseline and the perturbation procedure, which must be documented. Stability can be tested by applying small, label-preserving changes and measuring similarity between explanation maps after alignment. Model-parameter randomisation checks can help determine whether the map is sensitive to learned parameters. These tests provide evidence about specific properties, not a complete measure of human usefulness or causal explanation.

Calibration and abstention: If the application uses confidence thresholds to send uncertain cases for review, assess calibration with expected calibration error and reliability diagrams, while noting their sensitivity to binning. Temperature scaling may be fitted on the validation set only. A simple abstention analysis can measure the fraction of examples referred for human review and the error rate among predictions retained for automatic routing. Thresholds should be chosen according to an explicit operational objective, not after observing the final test results.

Efficiency evaluation: Record model parameter count, checkpoint size, inference latency, and, if available, memory use. Measure latency after warm-up, over a predefined number of repeated inferences, using the same device and batch size. Report median and dispersion as well as hardware, framework version, precision mode, and whether preprocessing is included. Avoid comparing GPU latency for one model with CPU latency for another. Energy use can be included if measured with a defensible instrument; otherwise, do not present parameter count as a direct measurement of energy consumption.

Statistical analysis: Use paired comparisons because each candidate model is evaluated on the same test images. Bootstrap confidence intervals can quantify uncertainty in macro-F1 or performance differences. For repeated training seeds, report between-seed variability. Multiple comparisons across many corruptions and classes increase the chance of false positives, so the analysis should emphasise effect sizes and confidence intervals and consider correction where formal hypothesis testing is used. The final analysis plan should be written before the test results are examined.

Reproducibility: Publish code, environment details, random seeds, model configurations, split identifiers where licence permits, and scripts for generating tables and figures. Do not redistribute copyrighted images unless the licence allows it; instead provide download instructions and preprocessing scripts. Document failed runs and deviations from the plan. A small repository with one-command evaluation is more useful than a polished demo whose reported numbers cannot be reproduced.

X. MEASUREMENT FRAMEWORK**Table 1. Proposed evaluation dimensions and measures**

Evaluation dimension	Measures	Purpose
Classification performance	Accuracy, macro-F1, per-class precision, per-class recall, and confusion matrix.	Establish how well the model assigns the intended label and reveal uneven performance across categories.
Robustness	Macro-F1 and recall under each corruption type and severity; absolute and relative change from clean performance.	Quantify sensitivity to common image degradation and identify conditions that merit review or additional training.
Explanation quality	Deletion/insertion behaviour, model-parameter randomisation checks, and explanation-map stability under label-preserving transformations.	Test selected properties of the explanation rather than treating visual appeal as proof of faithfulness.
Uncertainty	Calibration curve, expected calibration error, and risk-versus-coverage analysis where confidence-based abstention is used.	Examine whether confidence can support cautious routing and human review.
Efficiency	Model size, parameter count, median inference latency, latency dispersion, and peak memory where available.	Estimate whether the model is practical for the intended deployment environment.
Governance and documentation	Data provenance, licence, class balance, privacy review, known limitations, and human escalation path.	Make the conditions of use and the limits of the evidence explicit.

The measures should be reported in separate columns or panels. A model should not be labelled “best” merely because it has the highest clean accuracy. The selection decision should depend on the intended operating conditions and minimum acceptable performance on critical classes. If the application values recall for dangerous road damage more highly than precision, that priority must be stated and justified before results are interpreted.

XI. ETHICAL CONSIDERATIONS

The study should use data lawfully and responsibly. Dataset licences must be checked before downloading, modifying, or redistributing images. If new images are collected from residents, the project should obtain appropriate consent and institutional guidance. Images may contain identifiable people, number plates, home entrances, or precise location data. The research should avoid collecting such information unless necessary and should use redaction or minimisation procedures where appropriate.

The system must not imply that automated classification replaces public officials. Model predictions should be treated as recommendations, especially when confidence is low or the input differs from training data. A human review process should be available for uncertain cases and for users who believe a report has been misclassified. The interface should avoid presenting confidence as a probability of real-world correctness unless calibration has been evaluated under relevant conditions.

Dataset imbalance may create uneven service quality. If some classes or locations are underrepresented, aggregate metrics can hide systematic errors. The study should report support by class and, if location metadata is lawfully available, consider external or subgroup evaluation. It should not infer demographic fairness from image-corruption tests. Fairness claims require relevant group definitions, lawful data, appropriate metrics, and a clear justification for the comparison.

The paper should follow academic integrity standards. Any results included in the final manuscript must be produced by the described experiments. Images, charts, tables, and code from other sources must be credited and used according to their licences. The author should distinguish between the proposed framework and completed empirical findings, disclose material assistance where required by institutional policy, and avoid submitting the same manuscript to incompatible venues simultaneously.

XII. EXPECTED RESULTS AND ANALYTICAL PLAN

Because the experiments have not been run as part of this manuscript, this section describes expected analytical patterns rather than reporting findings. The first output will be a clean-test comparison showing whether candidate models differ in macro-F1, class-specific recall, and confusion patterns. A transfer-learning model may outperform a small baseline when the dataset is limited, but this is not guaranteed; data leakage, class imbalance, annotation noise, and domain mismatch can alter the result. The empirical conclusion must follow the measured evidence.

The corruption analysis is expected to reveal that not all image degradations have the same effect. Blur may remove fine details needed to distinguish shallow road damage from shadows, while low illumination may reduce visibility of drainage or waste features. Compression may have a smaller or larger impact depending on resolution and architecture. These are plausible mechanisms to investigate, not predetermined outcomes. The paper should report the actual degradation for each corruption and severity and avoid generalising beyond the tested conditions.

The explanation analysis will examine whether highlighted regions overlap with visible issue regions and whether removing highly ranked pixels affects model confidence as expected. If explanations remain visually similar after model parameters are randomised, that would raise concerns about whether they reflect learned decision behaviour. If explanations change dramatically under small, label-preserving transformations, their reliability for human review may be limited. The results should include both representative images and quantitative summaries to avoid cherry-picking compelling examples.

The efficiency analysis may reveal a trade-off between model capacity and deployment cost. A compact model could be attractive for low-cost hosting, while a larger model may provide stronger performance on some classes. The decision should not

be made on model size alone; latency, memory, throughput, and robustness are also relevant. If a model offers only a negligible performance improvement at a much greater cost, the simpler alternative may be preferable for a student prototype or resource-constrained public service.

The final result should be a transparent scorecard. For each model, report clean metrics, corruption metrics, explanation measures, calibration where evaluated, and latency. Include a failure-case analysis and a list of conditions under which the system should abstain. Do not create a single weighted ranking unless weights are explicitly justified with stakeholder input or a sensitivity analysis. The intended outcome is a decision-support framework, not a claim that one universal model is optimal.

XIII. DISCUSSION

The proposed evaluation reframes civic image classification as a reliability problem rather than a competition for the highest benchmark score. This is important because the application is embedded in a workflow where incorrect routing can consume staff time and delay attention to a local problem. A model that performs well on average but fails on a rare class may be unsuitable even if its overall accuracy is high. Macro-averaged and per-class measures make these weaknesses more visible.

Robustness evaluation provides a structured way to test foreseeable variation in citizen-submitted images. The most useful output is not simply a corruption leaderboard; it is an understanding of which conditions cause failure and whether the system can detect them. If performance falls sharply under low light, the application might request a clearer photograph, route the case for manual review, or use a different data collection process. Such mitigations should be tested rather than assumed effective.

Explainability should be interpreted with caution. Visual maps can help developers inspect whether the model attends to the issue or to irrelevant background features. Yet maps are not self-validating evidence. Quantitative perturbation checks and randomisation tests add useful safeguards, while user studies would be required to establish whether explanations help public employees make better decisions. The proposed paper therefore avoids the stronger claim that a heatmap makes the system transparent in every meaningful sense.

A combined scorecard can support a more defensible selection decision. In an early prototype, clean accuracy and latency may be central; in a system used across varied weather and camera conditions, robustness may matter more. In a human-review interface, calibrated uncertainty and explanation stability may influence whether staff can appropriately rely on recommendations. These priorities should be defined with intended users and revisited as the system moves from laboratory testing to field deployment.

The study also has implications for student research practice. A credible contribution does not require inventing a new architecture. It can arise from a carefully designed comparison, a clear research gap, reproducible implementation, honest treatment of uncertainty, and a meaningful analysis of limitations. The strongest paper will show that the author understands why metrics were selected, how leakage was prevented, why explanations were tested, and what the results do and do not support.

XIV. LIMITATIONS

Several limitations are anticipated. First, available datasets may not represent the full variety of civic issues, local conditions, or image quality found in real public reporting. A dataset focused on road damage cannot support conclusions about waste, drainage, and streetlight faults without additional evidence. Second, class labels may be noisy or inconsistent, and the original data collection process may introduce geographic or device-related bias.

Third, synthetic corruptions are proxies for real-world variation. They provide control and reproducibility but may not capture combinations of weather, motion, lighting, and scene complexity. Fourth, explanation metrics measure selected properties and do not fully capture human interpretability or causal relevance. Fifth, latency measured on one device may not transfer to another deployment environment. These limitations should be reported alongside the findings rather than relegated to a brief disclaimer.

Finally, a technical evaluation does not establish social impact. The study will not show that automated classification reduces complaint-resolution time unless a real workflow is evaluated. Future work should include prospective field testing, human factors evaluation, external validation across locations, privacy review, and monitoring for changes in data distribution after deployment.

XV. CONCLUSION

This paper proposes a practical framework for evaluating AI-based civic-issue classification beyond accuracy. It integrates clean-test performance, class-specific errors, robustness to controlled image corruptions, quantitative explanation checks, uncertainty analysis, and inference efficiency. The framework is motivated by the needs of civic reporting systems, where a model prediction can influence how a complaint is routed and how quickly it receives attention.

The central argument is that no single metric can establish readiness for deployment. Accuracy does not show how a model behaves under degraded images; a visually convincing heatmap does not prove explanation faithfulness; and a strong test score does not establish that the model generalises to another city or dataset. A transparent evaluation should expose these dimensions and report limitations clearly.

The next step is empirical implementation using a suitably licensed and documented dataset, a predefined experiment plan, and reproducible code. Once experiments have been completed, the manuscript should be revised to include actual results, confidence intervals, tables, figures, and a discussion grounded in those results. Until then, the proposed methods and expected patterns should not be represented as completed findings. This disciplined approach can produce a technically credible research artifact for further review and potentially support future work on responsible AI for public services.

XVI. REFERENCES

- [1] Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I., Hardt, M., & Kim, B. (2018). Sanity checks for saliency maps. *Advances in Neural Information Processing Systems*, 31.
- [2] Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*.
- [3] Geirhos, R., Jacobsen, J.-H., Michaelis, C., Zemel, R., Brendel, W., Bethge, M., & Wichmann, F. A. (2020). Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2, 665–673.
- [4] Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daumé III, H., & Crawford, K. (2021). Datasheets for datasets. *Communications of the ACM*, 64(12), 86–92.
- [5] Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017). On calibration of modern neural networks. *Proceedings of the 34th International Conference on Machine Learning*, 1321–1330.
- [6] Hendrycks, D., & Dietterich, T. (2019). Benchmarking neural network robustness to common corruptions and perturbations. *International Conference on Learning Representations*.
- [7] Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 30.
- [8] Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D., & Gebru, T. (2019). Model cards for model reporting. *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 220–229.
- [9] Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). “Why should I trust you?” Explaining the predictions of any classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1135–1144.
- [10] Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., & Fei-Fei, L. (2015). ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision*, 115, 211–252.
- [11] Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-CAM: Visual explanations from deep networks via gradient-based localization. *Proceedings of the IEEE International Conference on Computer Vision*, 618–626.
- [12] Sagawa, S., Koh, P. W., Hashimoto, T. B., & Liang, P. (2020). Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. *International Conference on Learning Representations*.
- [13] National Payments Corporation of India is not a source for this computer-vision study and is therefore not cited. Dataset-specific citations and licence information must be added after the final dataset has been selected and verified.