Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 3, Issue 10 (2025)

Online ISSN 3007-3197
Print ISSN 3007-3189

http://amresearchreview.com/index.php/Journal/about

Metamorphic Testing Framework for Deep Learning-Based Smart Traffic Signal Management

Muhammad Iqbal*
Department of Computer Science, University of Baltistan Skardu, Gilgit Baltistan, 16100, Pakistan Email: muhammad.iqbal@uobs.edu.pk

Jawad Usman Arshed
Department of Computer Science, University of Baltistan Skardu, Gilgit Baltistan, 16100, Pakistan

Ghulam Hussain
Department of Computer Science, University of Baltistan Skardu, Gilgit Baltistan, 16100, Pakistan

Dilawar Abbas
Department of Computer Science, University of Baltistan Skardu, Gilgit Baltistan, 16100, Pakistan

Muhammad Asim
Department of Computer Science, University of Baltistan Skardu, Gilgit Baltistan, 16100, Pakistan

Mudassir Ali
Department of Computer Science, University of Baltistan Skardu, Gilgit Baltistan, 16100, Pakistan

ABSTRACT

Deep learning models are central to modern in- telligent traffic systems but remain difficult to validate due to the absence of explicit test oracles. This study presents a Metamorphic Exploration (ME) framework to systematically evaluate the robustness and reliability of deep learning models in real-world traffic scenarios. The approach is demonstrated using the YOLO (You Only Look Once) model within a Smart Traffic Signal Management System. Metamorphic Relations (MRs) are defined based on key testing properties—robustness and effec- tiveness—using metrics such as accuracy, precision, recall, and class loss. Experimental results on both public and real-world datasets show that ME effectively measures model reliability and performance consistency under diverse traffic conditions.

Index Terms—Metamorphic Exploration, Deep Learning, YOLO, Traffic Signal Management, Model Robustness

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 3, Issue 10 (2025)

Online ISSN 3007-3197
Print ISSN 3007-3189

http://amresearchreview.com/index.php/Journal/about

INTRODUCTION

Regardless of machine learning's (ML) success in traffic control, a significant challenge remains: guaranteeing that these models operate consistently in a variety of scenarios. The lack of a comprehensive way to confirm that ML models are right for all potential inputs For instance, there is no reliable method to verify whether a model's vehicle count in a smart traffic system is consistently correct in situations such as heavy rainfall, dim lighting, or unusual vehicle kinds. Traditional testing approaches rely on labeled datasets or simulations, but these fail to expose hidden information when the model encounters edge cases (e.g., obscured vehicles or sensor er- rors).These hidden failure modes are dangerous in autonomous systems, where incorrect decisions like miscounting vehicle scan result in traffic mismanagement or even accidents. Thus, there is an urgent need for innovative validation techniques that transcend conventional methods to ensure robustness without requiring explicit "correct answers" for every scenario.

Smart intelligent systems are revolutionizing urban infras- tructure, with machine learning (ML) at the forefront of this transformation. In real-time applications like traffic signal con- trol, ML models enable dynamic decision-making that adapts to changing conditions. Traffic jams remains a critical urban challenge, causing significant delays, increased pollution, and

*All correspondence should be addressed to Dr. Muhammad Iqbal

excessive fuel consumption. For instance, studies show that traffic jams cost cities billion annually in lost productivity and environmental damage [5]. To address this, intelligent traffic management systems used machine learning models, particularly object detection models like YOLO (You Only Look Once), to analyze live video feeds, count vehicles, and optimize signal timings. These systems promise smoother traffic flow, reduced wait times, and safer roads [10]. However, their commercial success depends on the reliability of the underlying ML models in unpredictable real-world scenarios [15].

It is also found that, unlike traditional sensor-based systems, YOLO can process high-resolution visual data with remarkable speed and accuracy, making it ideal for urban environments with complex and dynamic traffic patterns [6], [10]. Addition- ally, these models allow for adaptive traffic control decisions based on actual conditions rather than fixed schedules, leading to smarter, more responsive city infrastructure. As such, ML is not just enhancing traffic systems it is reshaping how cities think about mobility and public safety [8]. The efficiency of a machine learning system refers to training and prediction speed. However, with the exponential increase in data volume and system complexity, efficiency can become more important than accuracy in certain real-time applications [17].

Despite the effectiveness of ML in traffic management, a major challenge remains, such as ensuring these models behave reliably under diverse conditions. Most ML models, including YOLO, suffer from the "oracle problem" due to the absence of a universal method to verify their correctness for every possible input. For example, in a smart traffic system, there is no foolproof way to confirm whether a model's vehicle count is always accurate across scenarios like heavy rain, low light, or unusual vehicle types [6], [15]. Traditional testing approaches rely on labeled datasets or simulations, but these fail to expose hidden flaws when the model encounters edge cases (e.g.,

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 3, Issue 10 (2025)

Online ISSN 3007-3197
Print ISSN 3007-3189
http://amresearchreview.com/index.php/Journal/about

obscured vehicles or sensor errors). In safety- critical applications like traffic control, such undetected errors can lead to system failures, congestion, or accidents. Thus, there is an urgent need for innovative validation techniques that transcend conventional methods to ensure robustness without
depending on manually labeled ground truth for every situation [15], [17].
Metamorphic testing (MT) [2], [3], [22] is one of the most popular approaches, alleviating the oracle problem and has been successfully adopted in different domains [35], [1], [4],
[**?**], [9], [25], [29], [31], [33]
Instead of focusing on the correctness of individual out- puts, MT examines the relations called metamorphic relations (MRs), among the multiple executions of the SUT, an MR violation typically indicates the existence of a bug [11], [35]. In this study, we propose an AI-based traffic signal management system that dynamically adapts signal timings in real time based on vehicle density extracted from live image analysis. The system is designed to operate at the edge level, enabling low-latency decision-making in resource-constrained urban environments. To ensure the model's reliability under varying conditions, we introduce a Metamorphic Exploration (ME)-based validation approach to assess the behavior of the underlying machine learning algorithms in real-world traffic scenarios.
The key contributions of this research are as follows:
We present the application of Metamorphic Exploration (ME) to validate machine learning model behavior under real-world variability systematically.
We design and implement a dynamic, edge-deployable traffic signal management system that adjusts signal tim- ings based on real-time vehicle detection.
We conduct comprehensive experiments using a hybrid dataset that combines publicly available data with real- world traffic scenarios, demonstrating the robustness and practical applicability of the proposed approach.
The rest of the paper is organized as follows: Section II introduces the steps invloved in evaluating the proposed ap- proach. Section III presents the experimental results with a brief discusions. Section IV reviews some related work, and Section V concludes the paper.

METHODOLOGY

This section explains the main steps involved in evaluating the approach executions.
System Design and Setup
To evaluate the proposed approach, we initially developed a hardware-integrated Smart Traffic Signal Management System for a four-lane junction. This system utilizes the YOLO — a deep learning model for real-time vehicle detection. Figure 1 illustrates the four designated lanes (L1, L2, L3, and L4), with the respective vehicle counts of 0, 3, 2, and 5.
To conduct the experiment, cameras were mounted on a pole near a traffic signal to capture a clear view of traffic across Lanes 1, 2, 3, and 4. These images were then processed using the YOLO (You Only Look Once) algorithm to detect and count the number of vehicles present. Based on the vehicle count, the traffic density for the corresponding side was calculated.

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 3, Issue 10 (2025)

Online ISSN 3007-3197
Print ISSN 3007-3189

http://amresearchreview.com/index.php/Journal/about

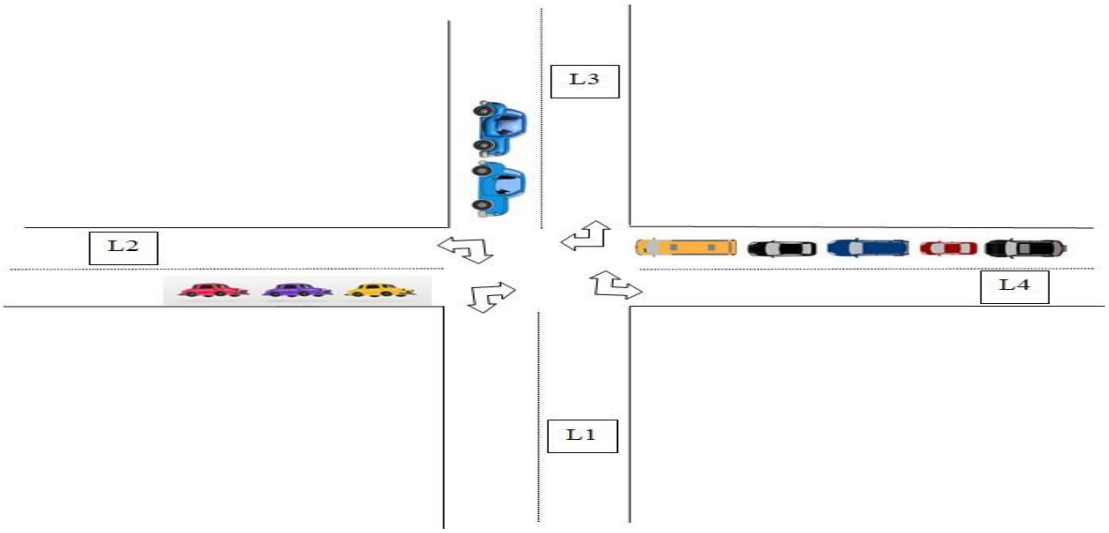

Fig. 1: **Overview of the four lanes traffic Junction.**

The system's workflow involves the following steps: Ini- tialization — activates cameras and begins data acquisition. Vehicle Detection — capture images using cameras installed at intersections and detect vehicles using YOLO. Traffic Signal Adjustment — dynamically allocate green light duration based on real-time vehicle counts. Cycle Monitoring — continuously monitor and adjust signals for all lanes in a loop.
Table I shows four lanes, each with a varying number of vehicles. Lane 1 (L1) is occupied by no vehicle; hence the green light remains inactive. Lane 2 (L2) accommodates three vehicles, permitting a six-second interval for the green light. However, Lane 3 (L3) remains two vehicles; hence, allowing four seconds for the green light to illuminate. Lastly, Lane 4 (L4) has five vehicles, granting a ten-second duration for the green light to be active. This cycle continues in repetition.This dynamic allocation ensures that lanes with more vehicles get more green light time, improving traffic flow and reducing unnecessary waiting times.

TABLE I: **Working strategy of the proposed system.**

Lane	Vehicle Count	Allocated timing (s)
L1	0	0
L2	3	6
L3	2	4
L4	5	10

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 3, Issue 10 (2025)

Online ISSN 3007-3197
Print ISSN 3007-3189

http://amresearchreview.com/index.php/Journal/about

Data Collection and Preprocessing

To evaluate the behavior of the proposed model and to con- duct the case study, three types of datasets were employed: a publicly available dataset from Kaggle containing car images, a custom-built dataset comprising cars and motorbikes, and a real-world dataset including various vehicles such as cars, buses, and trucks. To ensure consistency and enhance model performance, all images underwent standardized preprocess- ing, including resizing to ensure input compatibility and data augmentation techniques such as rotation, flipping, and scaling to improve dataset diversity and reduce the risk of overfitting. Selected examples from these datasets are presented in Fig- ure 2.

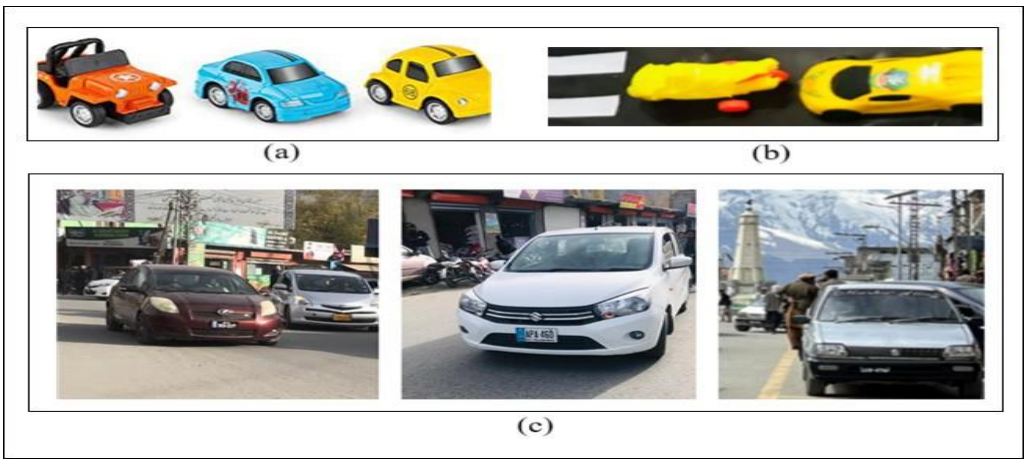

Fig. 2: Sample images from the three datasets used in this study: (a) cars from the Kaggle dataset, (b) motorbikes and cars from the custom dataset, and (c) real-world vehicle images from three different places at Skardu city.

To evaluate the model's effectiveness through Metamorphic Exploration (ME), we employed a combination of publicly available, custom, and real-world datasets. Initially, a pub- licly available dataset consisting of 1000 annotated vehicle images was used. From this dataset, 800 images were used for model training as the primary source input, while the remaining 200 images were reserved for validation during follow-up evaluations. In the second phase, a custom dataset comprising 320 motorbike images was used, where 250 images supported further training and 70 were employed in subsequent ME executions to validate the model's behavior under input variation.

To assess the model's generalizability and practical perfor- mance, we further incorporated a real-world dataset collected from high-traffic areas in Skardu City, specifically Yadgar Chowk, Alamdar Chowk, and Dr. Irshad Chowk. This dataset includes 200 images capturing diverse vehicles such as trucks, buses, and vans. It was used in the cross-validation phase to explore the model's robustness and reliability under real environmental conditions. A detailed summary of the datasets used is presented in Table II.

TABLE II: **Summary of Training and Testing Data set.**

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 3, Issue 10 (2025)

Online ISSN 3007-3197
Print ISSN 3007-3189

http://amresearchreview.com/index.php/Journal/about

Data Execution Types	Training (Source Executions)	Validation (Follow-up Executions)	Total Executions	Images Sources
Car	800	200	1000	Kaggle
Car & Bike	250	70	320	Custom
Car	160	40	200	Real Environment
Total	1210	310	1520	3

Metamorphic Relations for Ideal Model Behavior

To evaluate the model's reliability and to design metamor- phic relations based on key performance metrics—such as accuracy, precision, recall, and class loss—we employed the Metamorphic Exploration (ME) approach. In this process, we adopted the previously established machine learning testing properties of robustness and effectiveness.

MR_1 – Accuracy Consistency:

If a model "M" trained on a publicly available dataset **"Dpub"** giving high accuracy **"A"** in a define epochs **"C"**, Then the same model trained under comparable conditions on a custom dataset **"Dcust"** should yield similar accuracy **"A"**, assuming data quality and class distribution are comparable.

MR_2 – Precision and Recall Stability:

Let E_1 and E_2 be two training epochs where E_2 is greater than E_1 , i.e., the model is trained for more epochs un- der configuration E_2 . Assume all other conditions remain constant, including the dataset, model architecture, optimizer settings, and batch size. Then, the model's behavior in terms of classification performance- Precision and Recall should remain stable.

MR_3 – Class Loss Behavior: With all training conditions held constant, increasing the number of epochs should result in class loss that either decreases or remains stable, indicating consistent learning.

Execution of Metamorphic Relations

During the execution of the identified metamorphic rela- tions, the deep learning model was trained using three distinct datasets across varying epoch ranges, serving as source test cases. The experimental results revealed a gradual improve- ment in accuracy, accompanied by stable precision and recall metrics, and a consistent class loss trend—demonstrating the robustness and reliability of the system under different training conditions.

MR_1 was executed to evaluate the model's behavior by initially training it with the publicly available dataset Dpub across defined epoch ranges. Subsequently, the custom dataset Dcust was used for follow-up executions, aiming to achieve comparable accuracy, denoted as A. During the initial train- ing phase with Dpub, the model demonstrated a consistent improvement in accuracy over increasing epoch ranges. For

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 3, Issue 10 (2025)

Online ISSN 3007-3197
Print ISSN 3007-3189
http://amresearchreview.com/index.php/Journal/about

instance, the accuracy improved from 86.3% in the epoch range 1–10 to 93.2% in the epoch range 41–50. These results are summarized in the last column of Table III.

TABLE III: **Accuracy Progression Across Epoch Ranges for MR_1 Evaluation.**

Epoch Range	Class Loss	Precision (P)	Recall (R)	Accuracy (%)
1-10	0.851	0.912	0.865	86.3
11-20	0.829	0.915	0.872	88.2
21-30	0.782	0.887	0.843	89.4
31-40	0.761	0.92	0.825	91.1
41-50	0.369	0.96	0.952	93.2

To evaluate the model's performance, we conducted follow- up executions using a real-world dataset to retrain model "M," while maintaining the same epoch range settings. During these runs, the accuracy initially dropped to 72.20% in the first epoch range (1–10), then unexpectedly increased to 80.72% in the second range (11–20), and gradually reached 93.77% by the epoch range 41–50. The sharp decline in accuracy from 86.3% to 72.20%, followed by a sudden rise to 80.72%, deviates from expected behavior and indicates a violation of MR_1. These results are summarized in the last column of Table IV.

TABLE IV: **Accuracy Trends and MR_1 Violation During Model "M" Retraining.**

Epoch Range	Class Loss	Precision (P)	Recall (R)	Accuracy (%)
1-10	1.783	0.805	0.655	72.20
11-20	1.037	0.924	0.715	80.72
21-30	0.883	0.907	0.864	88.53
31-40	0.725	0.958	0.934	94.63
41-50	0.699	0.949	0.926	93.77

To assess the model's classification behavior, Metamorphic Relations: MR_2 and MR_3 were applied, focusing on key performance metrics including precision, recall, and class loss. According to the metamorphic expectations, these metrics should remain relatively

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 3, Issue 10 (2025)

Online ISSN 3007-3197
Print ISSN 3007-3189

http://amresearchreview.com/index.php/Journal/about

stable during follow-up executions.

Our observations reveal that precision and recall remained con- sistent across the initial two epoch ranges (1–10 and 11–20). However, a noticeable deviation was observed in epoch range 21–30, where precision dropped to 0.88 before recovering to 0.92 in the subsequent range. Similarly, recall fluctuated from 0.84 to 0.82, then sharply increased to 0.95 over the final three epoch ranges. Furthermore, class loss exhibited an abrupt decline from 0.76 to 0.369 in epoch range 41–50. These unexpected variations in performance indicate violations of MR_2 and MR_3 , as they deviate from anticipated stability under Metamorphic Exploration (ME). A more detailed analysis of these results is provided in the following section.

RESULTS AND DISCUSSION

When investigating the MRs Violations, we found that, the Loss value — class loss decrease across epoch ranges, suggesting the model is learning effectively and becoming better at detecting and classifying objects. However the sudden fluctuation indicates the model need to learn more precisely to indicate better performance to handle the hidden oracles.

Precision (P) is consistently high (around 0.9), indicating that the model is performing well. There is a sharp drop in class loss around epoch range 41-50 (from 0.761 to 0.369), which could indicate that the model has learned to classify objects more confidently by this stage.

The observed fluctuation in YOLO model accuracy during follow-up training is primarily attributed to a domain shift between the original and real-world datasets, causing the model to initially struggle with unfamiliar data distributions. Additional contributing factors include suboptimal fine-tuning configurations (e.g., learning rate and layer freezing), class im- balance in the new dataset, and training instabilities introduced by small batch sizes or aggressive data augmentation. These issues collectively led to temporary performance degradation, indicating a violation of MR_1.

Figure 3 illustrates the model's performance trends across key metrics—accuracy, precision, recall, and class loss—highlighting deviations that signal violations of the defined Metamorphic Relations (MRs). Figure (a) presents the model's performance during execution on the publicly available dataset, while Figure (b) shows the corresponding performance when evaluated on the real-world dataset.

To avoid such violations in the future, it is essential to ensure proper alignment between the source and target datasets through preprocessing techniques that minimize domain shifts, such as normalization, augmentation harmonization, or domain adaptation methods. Additionally, maintaining a balanced class distribution in training batches and selecting stable batch sizes can help ensure consistent learning behavior ensuring the satisfaction to expected metamorphic behavior.

The violations of MR_2 and MR_3 observed in the fluctuating precision, recall, and class loss are primarily caused by train- ing instability during follow-up executions. Key contributing factors include inconsistent data distribution, particularly class imbalance and varying sample difficulty across batches, which

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 3, Issue 10 (2025)

Online ISSN 3007-3197
Print ISSN 3007-3189

http://amresearchreview.com/index.php/Journal/about

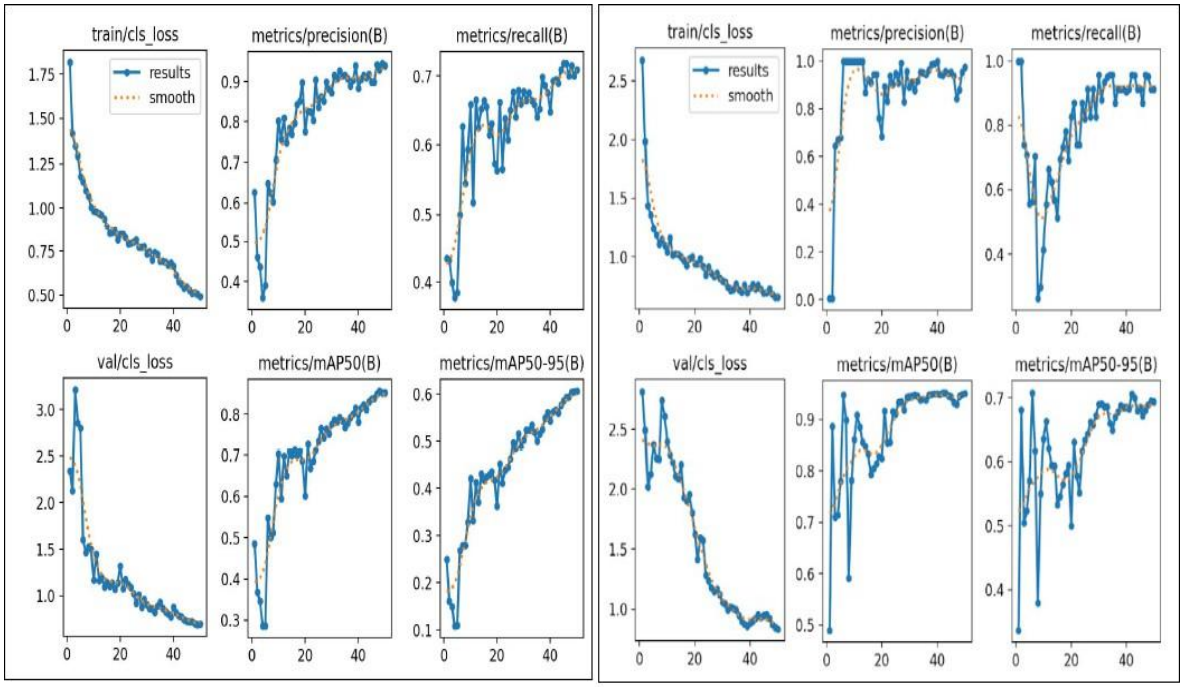

Fig. 3: Performance trends of the model across accuracy, precision, recall, and class loss metrics: (a) execution on the public dataset; (b) execution on the real-world dataset, highlighting MR violations.

can lead to temporary performance drops. Inadequate fine- tuning strategies—such as improper learning rate adjustment or failure to stabilize layers like batch normalization—can fur- ther exacerbate metric fluctuations. These factors collectively prevent the model from maintaining the expected stability in classification behavior, thus violating the metamorphic expectations.

To prevent violations of MR_2 and MR_3 in future executions, it is essential to implement a more controlled and stable training pipeline. This includes ensuring class balance across training batches to avoid biased learning, using a consistent and representative sampling strategy. Additionally, limiting overly aggressive data augmentations and monitoring valida- tion metrics at finer epoch intervals can help detect and correct deviations early, ensuring the model adheres to the expected metamorphic behavior.

Overall, we applied ME as validation techniques by feeding the remaining subsets of data into the trained model. During this phase, certain epochs exhibited inconsistent behavior, revealing fluctuations in model performance that deviated from expected outcomes. These anomalies highlight the relevance of metamorphic exploration, emphasizing its effectiveness in uncovering hidden weaknesses and guiding areas for model improvement in future implementations.

RELATED WORKS

Metamorphic Testing (MT) has emerged as a valuable tech- nique for validating the reliability and robustness of machine learning (ML) systems, particularly in scenarios where the

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 3, Issue 10 (2025)

Online ISSN 3007-3197
Print ISSN 3007-3189
http://amresearchreview.com/index.php/Journal/about

absence of test oracles limits the effectiveness of traditional testing approaches [11], [23], [34]. Several studies have ex- plored its applicability across diverse domains, including au- tonomous driving [12], [13], hydrology [21], [30], generative models [19], and deep learning infrastructure [28].

Fu et al. introduced MetaFL, a weakly-supervised fault localization framework that combines MT with code cover- age analysis. A key innovation was the MTG (Metamorphic Test Grouping) technique, which clusters related tests to enhance fault detection. MetaFL effectively localized faults in ML code components without requiring detailed fault labels, demonstrating improvements in both accuracy and efficiency compared to traditional debugging methods [7].

A ModelMeta —model-level metamorphic testing technique
— was designed by Mu et al. to enhance the testing of deep learning frameworks such as PyTorch, MindSpore, and ONNX Utilizing four structure-based metamorphic relations and a QR-DQN reinforcement learning strategy, ModelMeta automatically generates diverse test models to uncover bugs by analyzing loss, gradients, memory, and execution time. Evaluated on 17 models across 10 tasks, it detected 31 new bugs (27 verified, 11 fixed) and achieved up to 28.7% code coverage and 78.6% API coverage, outperforming existing methods in precision and efficiency [18].

In the hydrological modeling domain, Reiniger et al. applied MT to both conceptual and machine learning-based rainfall- runoff models. The authors defined MRs simulating realistic input shifts—such as rainfall intensity and duration—and
analyzed model responses. Their findings revealed behavioral inconsistencies between model types that standard perfor- mance metrics failed to capture, thereby reinforcing MT's role in enhancing model robustness for environmental decision- making systems [21].

A recent contribution in 2025 introduced an AI-augmented MT approach for autonomous vehicle testing. The framework uses AI to generate diverse scenario variations, such as altered lane markings and weather conditions, to define MRs and sig- nificantly expand testing coverage. This automated approach uncovered critical failures in perception, planning, and control modules that conventional tests overlooked, demonstrating the potential of AI-augmented MT in safety-critical domains [32]. Also in 2025, a statistical metamorphic testing (SMT) framework was proposed for CNN-based deep learning sys- tems, particularly in medical imaging, where output random- ness is common. By incorporating statistical analysis to eval- uate consistency in output distributions under transformations such as brightness or orientation shifts, the study identified 85.7% of real faults—substantially outperforming traditional
deterministic testing methods [20].

Mendez et al introduces a hybrid technique that integrates Metamorphic Testing (MT) with Machine Learning (ML) to improve data quality in OpenStreetMap (OSM). Using glob- ally available map data, Random Forest models automatically filter false positives from MT outputs, achieving over 90% accuracy in detecting genuine tagging errors—surpassing the performance of MT or ML alone. The proposed SMT frame- work significantly reduces manual verification and enhances the reliability of large-scale mapping systems [16].

Huang et al. proposed MT-YOLO, a framework for evalu- ating YOLOv3's reliability under controlled transformations, including brightness changes, object occlusion, and

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 3, Issue 10 (2025)

Online ISSN 3007-3197
Print ISSN 3007-3189
http://amresearchreview.com/index.php/Journal/about

geometric rotations. The authors used predefined metamorphic relations (MRs) to transform input images and observed that YOLO's object detection output was consistent. Their results high- lighted that YOLOv3 frequently failed to maintain consistent bounding boxes or class probabilities under seemingly minor visual alterations, exposing significant vulnerabilities in real- world applications.

Similarly, Chien et al. applied MT to YOLOv4-based traffic sign recognition systems under realistic driving conditions, such as rain effects, motion blur, and camera angle distortions. The MRs used were domain-specific, ensuring that detected traffic signs would remain consistent despite such variations. Despite YOLO's high accuracy in ideal conditions, the model's performance degraded significantly with even slight weather or motion-induced distortions, raising concerns about its reli- ability in autonomous driving environments [26].

Wang et al. extended MT to medical object detection tasks using YOLOv5 and integrated a statistical metamor- phic testing (SMT) framework. Unlike traditional MT, SMT quantitatively assessed output distributions to detect deviations across transformations such as orientation shifts and added Gaussian noise. This statistical rigor helped uncover subtle inconsistencies in bounding box placements and detection probabilities, especially in low-contrast regions of medical images, where diagnostic sensitivity is critical [27].

Bingyi Su [24] investigated the use of Metamorphic Testing (MT) to verify deep learning–based image recognition sys- tems, addressing the oracle problem of undefined expected outputs. Using the ResNet-50 model trained on ImageNet-1K within the GeMTest framework, the study demonstrated MT's capability to detect classification errors but also revealed a high false-positive rate, emphasizing the need for greater accuracy and automation in future AI testing methods.

Recently, Iqbal et al. proposed a user-centered approach that integrates metamorphic relations (MRs) into the Soft- ware Development Life Cycle (SDLC) to improve the quality and evaluation of Learning Management Systems (LMS). By aligning MRs with different SDLC phases, the study enhances the accuracy and reliability of user feedback and evaluation results, leading to more informed decisions and improved LMS development and refinement. However, having the future potential to automate the MR validation process using intelligent algorithms or AI-based algorithms [14]

Conclusion and Future Work

In this study, we applied Metamorphic Exploration (ME) to evaluate the reliability and classification behavior of the YOLO model within a Smart Traffic Signal Management System. By designing and executing a set of Metamorphic Relations (MR_1–MR_3), we uncovered significant insights into the model's performance under varying data conditions. The analysis revealed that although the model demonstrated consis- tent learning trends—evident in the decreasing class loss and generally high precision—several fluctuations were observed across accuracy, recall, and loss metrics, particularly in the follow-up executions on real-world datasets. These deviations indicate violations of the defined MRs, primarily caused by domain shifts, training instabilities, class imbalances, and suboptimal fine-tuning practices.

Our findings affirm the utility of ME in exposing such hidden behaviors that traditional validation technique may overlook. By visualizing the trends across different metrics,

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 3, Issue 10 (2025)

Online ISSN 3007-3197
Print ISSN 3007-3189
http://amresearchreview.com/index.php/Journal/about

we highlighted the model's sensitivity to data distribution and training configuration, emphasizing the need for more robust validation frameworks when deploying deep learning models in real-world, safety-critical applications.

In the future, we aim to expand the ME framework by incorporating automated detection and classification of MR violations using statistical and learning-based techniques. Conducting cross-validation on additional deep learning mod- els such as COCO, Faster R-CNN, and SSD (Single Shot Detector) represents a valuable and promising direction for future research. Additionally, we plan to explore adaptive fine-tuning methods, domain adaptation strategies, and meta- learning approaches to enhance model generalizability and stability. Applying ME to a broader range of object detection models and real-time systems will further validate its scalabil- ity and effectiveness as a testing paradigm for deep learning- based intelligent systems.

DATA AVAILABLITY

The dataset is available for further research and can be downloaded by interested individuals from: https://www.kaggle.com/datasets/hanif535/real-environment- vehicle-images."

ACKNOWLEDGMENTS

We wish to thank the open source community for proving the data. Special thanks to Mr hanif ,Asadi and Raziq for their support in collecting the real data from our surroundings.

REFERENCES

J. Brown, Z. Q. Zhou, and Y.-W. Chow, "Metamorphic testing of navigation software: A pilot study with Google Maps," in Proceed- ings of the 51st Annual Hawaii International Conference on System Sciences (HICSS-51), 2018, pp. 5687–5696, available: http://hdl.handle. net/10125/50602.

T. Y. Chen, T. H. Tse, and Z. Q. Zhou, "Fault-based testing without the need of oracles," Information and Software Technology, vol. 45, no. 1,
pp. 1–9, 2003.

T. Y. Chen, F.-C. Kuo, H. Liu, P.-L. Poon, D. Towey, T. H. Tse, and Z. Q. Zhou, "Metamorphic testing: A review of challenges and opportunities," ACM Computing Surveys, vol. 51, no. 1, pp. 4:1–4:27, 2018.

T. Y. Chen, F.-C. Kuo, W. Ma, W. Susilo, D. Towey, J. Voas, and Z. Q. Zhou, "Metamorphic testing for cybersecurity," Computer, vol. 49, no. 6,
pp. 48–55, 2016.

Y. Deng, X. Zheng, M. Zhang, G. Lou, and T. Zhang, "Scenario-based test reduction and prioritization for multi-module autonomous driving systems," in Proceedings of the 30th ACM Joint European Software Engineering Conference and Symposium on the Foundations of Software Engineering, 2022, pp. 82–93.

M. Flores-Calero, C. A. Astudillo, D. Guevara, J. Maza, B. S. Lita,
B. Defaz, J. S. Ante, D. Zabala-Blanco, and J. M. Armingol Moreno, "Traffic sign detection and recognition using yolo object detection algorithm: A systematic review," Mathematics, vol. 12, no. 2, p. 297, 2024.

L. Fu, Y. Lei, M. Yan, L. Xu, Z. Xu, and X. Zhang, "Metafl: Meta- morphic fault

Online ISSN 3007-3197
Print ISSN 3007-3189
http://amresearchreview.com/index.php/Journal/about

localisation using weakly supervised deep learning," IET Software, vol. 17, no. 2, pp. 137–153, 2023.

J. Gao, Y. Shen, J. Liu, M. Ito, and N. Shiratori, "Adaptive traffic signal control: Deep reinforcement learning algorithm with experience replay and target network," arXiv preprint arXiv:1705.02755, 2017.

J. C. Han and Z. Q. Zhou, "Metamorphic fuzz testing of autonomous ve- hicles," in Proceedings of the IEEE/ACM 42nd International Conference on Software Engineering Workshops (ICSEW '20). ACM, 2020.

A. Hazarika, N. Choudhury, M. M. Nasralla, S. B. A. Khattak, and I. U. Rehman, "Edge ml technique for smart traffic management in intelligent transportation systems," IEEE Access, vol. 12, pp. 25 443–25 458, 2024.

M. Iqbal, "Metamorphic Relations for Testing Automated and Au- tonomous Driving Systems and Simulation Platforms," Ph.D. disserta- tion, University of Wollongong, 2024.

M. Iqbal, J. C. Han, Z. Q. Zhou, and D. Towey, "Enhancing Euro NCAP Standards with Metamorphic Testing for Verification of Advanced Driver-Assistance Systems," in 2021 IEEE/ACM 6th International Work- shop on Metamorphic Testing (MET). IEEE, 2021, pp. 37–41.

M. Iqbal, J. C. Han, Z. Q. Zhou, D. Towey, and T. Y. Chen, "Metamor- phic testing of advanced driver-assistance system (ADAS) simulation platforms: lane keeping assist system (LKAS) case studies," Information and Software Technology, vol. 155, p. 107104, 2023.

M. Iqbal, G. Hussain, J. U. Arshed, B. Ali, D. Hussain, and Y. Hussain, "User-centric metamorphic relations: Enhancing learning management systems across development phases," pp. 1–6, 2025.

B. Ko″va´ri, L. Szo″ke, T. Be´csi, S. Aradi, and P. Ga´spa´r, "Traffic signal control via reinforcement learning for reducing global vehicle emission," Sustainability, vol. 13, no. 20, p. 11254, 2021.

M. Me´ndez, A. Becerra-Tero´n, J. M. Almendros-Jime´nez, M. G. Mer- ayo, and M. Nu´n˜ez, "Combining metamorphic testing and machine learning to enhance openstreetmap," IEEE Transactions on Reliability, vol. 73, no. 4, pp. 1834–1848, 2024.

C. Mohankumar and A. Manikandan, "Decentralized traffic management with federated edge ai: a reinforced transnet model for real-time vehicle object detection and collaborative route optimization," Discover Applied Sciences, vol. 7, no. 7, p. 729, 2025.

Y. Mu, J. Zhai, C. Fang, X. Chen, Z. Cao, P. Yang, K. Zhao, A. Guo, and
Z. Chen, "Improving Deep Learning Framework Testing with Model- Level Metamorphic Testing," Proceedings of the ACM on Software Engineering, vol. 2, no. ISSTA, pp. 2158–2180, 2025.

H. Park, T. Waseem, W. Q. Teo, Y. H. Low, M. K. Lim, and C. Y. Chong, "Robustness evaluation of stacked generative adversarial networks using metamorphic testing," in 2021 IEEE/ACM 6th International Workshop on Metamorphic Testing (MET). IEEE, 2021, pp. 1–8.

F. U. Rehman and C. Izurieta, "Testing convolutional neural network based deep learning systems: a statistical metamorphic approach," PeerJ. Computer science,

Online ISSN 3007-3197
Print ISSN 3007-3189

vol. 11, p. 2658, 2025.

P. Reichert, K. Ma, M. Ho¨ge, F. Fenicia, M. Baity-Jesi, D. Feng, and C. Shen, "Metamorphic testing of machine learning and conceptual hydrologic models," Hydrology and Earth System Sciences, vol. 28, no. 11, pp. 2505–2529, 2024.

S. Segura, G. Fraser, A. B. Sanchez, and A. Ruiz-Corte´s, "A survey on metamorphic testing," IEEE Transactions on Software Engineering, vol. 42, no. 9, pp. 805–824, 2016.

S. Segura, D. Towey, Z. Q. Zhou, and T. Y. Chen, "Metamorphic testing: Testing the untestable," IEEE Software, in press (accepted on 9 August 2018).

B. Su, "A comparative study of the application of metamorphic testing in image recognition and processing," 2025.

Y. Tian, K. Pei, S. Jana, and B. Ray, "DeepTest: Automated testing of deep-neural-network-driven autonomous cars," in Proceedings of the IEEE/ACM 40th International Conference on Software Engineering (ICSE '18). ACM, 2018, pp. 303–314.

C. Wang, B. Zheng, and C. Li, "Efficient traffic sign recognition using yolo for intelligent transport systems," Scientific Reports, vol. 15, no. 1,
p. 13657, 2025.

S. Wang and Z. Su, "Metamorphic testing for object detection systems,"
arXiv preprint arXiv:1912.12162, 2019.

D. Xiao, Z. Liu, Y. Yuan, Q. Pang, and S. Wang, "Metamorphic testing of deep learning compilers," Proceedings of the ACM on Measurement and Analysis of Computing Systems, vol. 6, no. 1, pp. 1–28, 2022.

Y. Xu, Z. Q. Zhou, X. Zhang, J. Wang, and M. Jiang, "Metamorphic testing of named entity recognition systems: A case study," IET Software,
pp. 386–404, 2022.

Y. Yang and T. F. M. Chui, "Reliability assessment of machine learning models in hydrological predictions through metamorphic testing," Water Resources Research, vol. 57, no. 9, p. e2020WR029471, 2021.

M. Zhang, Y. Zhang, L. Zhang, C. Liu, and S. Khurshid, "DeepRoad: GAN-based metamorphic testing and input validation framework for autonomous driving systems," in Proceedings of the 33rd ACM/IEEE International Conference on Automated Software Engineering (ASE '18). ACM, 2018, pp. 132–142.

T. Zhang, B. Kantarci, and U. Siddique, "Ai-augmented metamorphic testing for comprehensive validation of autonomous vehicles," in 2025 IEEE/ACM 1st International Workshop on Software Engineering for Autonomous Driving Systems (SE4ADS). IEEE, 2025, pp. 1–4.

Z. Q. Zhou and L. Sun, "Metamorphic testing of driverless cars,"
Communications of the ACM, vol. 62, no. 3, pp. 61–67, March 2019.

Z. Q. Zhou, L. Sun, T. Y. Chen, and D. Towey, "Metamorphic relations for enhancing system understanding and use," IEEE Transactions on Software Engineering, vol. 46, no. 10, pp. 1120–1154, 2018.

Z. Q. Zhou, S. Xiang, and T. Y. Chen, "Metamorphic testing for software quality assessment: A study of search engines," IEEE Transactions on Software Engineering, vol. 42, no. 3, pp. 264–284, 2016.

Annual Methodological Archive Research Review
http://amresearchreview.com/index.php/Journal/about
Volume 3, Issue 10 (2025)

Online ISSN 3007-3197
Print ISSN 3007-3189

http://amresearchreview.com/index.php/Journal/about