OXFORD JOURNAL
OF STUDENT
SCHOLARSHIP

Impact of Dataset Diversity on Traffic Sign Detection Models Under Harsh Visual Conditions

Jonah Lee
jonahlee1189@gmail.com

ABSTRACT

Object detection models represent a deeply important element of autonomous vehicles. These systems rely on computer vision in order to detect, classify, and respond to important road marks. In particular, traffic sign detection models are known to be highly accurate, yet their performance continues to face limitations when exposed to harsh visual conditions such as fog, poor lighting, or snow. While dataset diversity remains a possible solution to addressing this problem, the field lacks quantitative studies which establish a clear relationship between the diversity of datasets and model performance.

This study aimed at identifying whether increasing the diversity of training datasets could improve the model accuracy for detecting signs in harsh visual conditions. All images were extracted from the Mapillary Traffic Sign Dataset, and assigned four environmental metrics: brightness, contrast, blur, and harshness. Then, images were grouped to construct three different training datasets of increasing diversity: LOW, MEDIUM, and HIGH. In order to quantify these levels, each dataset was assigned a Vendi score to calculate a numeric diversity measurement based on the different image characteristics. In addition, a HARSH test dataset was constructed by taking the photos with the most visually challenging conditions. Three object detection models, YOLOv5s, YOLOv7-tiny, and YOLOv8n were tested on the three training datasets, then all evaluated on the HARSH set, measuring their performance using precision, recall, F1-score, and mAP@50.

Results demonstrated no significant or consistent correlation, refuting the initial hypothesis that there would be a clear relationship leading to an increase in accuracy. These findings demonstrate diversity alone is not sufficient enough to improve model robustness, as dataset size and model architecture may play a more significant role.

INTRODUCTION

As autonomous vehicles become increasingly prominent, their dependability and safety on the roads remains a key focus. Among the operations of these vehicles is object detection models, which play an important role in how these cars are able to see and respond to their surroundings. Traffic sign detection is

particularly important because it is vital for them to be able to identify, classify, and interpret essential markings which determine the safety of many others on the roads. While traffic sign detection models have demonstrated high accuracy under controlled conditions, their performance is greatly hindered by harsh visual conditions, such as being subjected to fog, bad lighting, or rain in real world environments (Contreras et al., 2024). These failures present a major challenge in the development and reliability of autonomous vehicles out in the field.

Addressing this issue, researchers have explored multiple approaches to improving model robustness. A widely discussed solution is increasing the environmental diversity of training datasets used to develop these models. In other wordings, feeding each computer a wider range of images with varying visual characteristics and conditions. The assumption is that exposing models to a broader scope during the training process will allow them to better acclimate to harsher and unexpected environments in real scenarios (Jadoun, 2023).

Despite the significant importance which it holds, dataset diversity remains largely unexplored within the field of traffic sign detection. Much of the existing literature relies on qualitative descriptions or observations to demonstrate the effects which modifying this aspect may have. However, a lack of controlled experimental design or quantitative benchmarking prevents findings from being isolated or completely attributed to a change in diversity. The
absence of a controlled, quantitative analysis creates uncertainty on the true impact of dataset diversity on the performance of traffic sign detection models.

To address this gap, this study investigates the relationship between dataset environmental diversity and object detection performance under harsh visual conditions. The independent variable, dataset diversity, is defined as the degree of variation in lighting, weather, and time-of-day conditions represented within a dataset, quantified using a diversity score calculated by using the proportions of different types of images in the dataset. The dependent variable, model performance, is defined as the detection accuracy of the traffic sign recognition models, using measurements such as precision, recall, F1 score, and mean average precision. The intervening variable, model architecture, is defined as the type of object detection model that will be used and controlled in order to ensure a clear comparison between the effects of the datasets. By isolating diversity as an independent variable and controlling for other factors, this study aims to determine whether increasing dataset diversity alone leads to measurable improvements in model robustness. The initial hypothesis was that higher levels of environmental diversity would reveal a significant, positive relationship with accuracy across multiple object detection models when evaluated under harsh conditions.

LITERATURE REVIEW

Convolutional Neural Networks and Object Detection

A majority of object detection models in autonomous vehicles are built on convolutional neural networks, or CNNs. CNNS are a class of machine learning architecture that learns to recognize visual patterns by repeatedly being shown labeled training images. During training, the model processes each image, compares its prediction against the correct label, and adjusts its internal parameters to reduce future errors (Youssouf, 2022). This cycle is repeated over and over across many images until the model is able to reliably make predictions for images that it has never even seen before. When analyzing a completely new image, CNNs extract information hierarchically. Early layers detect simple patterns like edges and pixel intensities, comparing this with the patterns similar to training images to determine whether or not a recognizable object is within the frame (Landge et al., 2025).

Among CNNs, the YOLO family of models has become one of the most dominant architecture types used for autonomous vehicles. YOLO processes the entire image in a single pass, simultaneously predicting object classes and locations (Qiu et al., 2025). This provides an efficient procedure for the hierarchy process mentioned by Landge et al. that is favored for being both accurate and quick under high speed, real life scenarios. In addition to optimized detection processes such as these, the training dataset used is crucial for the performance of these models. If training images reflect a narrow range of visual characteristics, then models will poorly perform when encountering environments outside of that range. Known as distribution shift, Mao et al. (2023) explains that the types of training images used almost directly translate into the generalizability and performance of object detection models out in the field. As a result, dataset composition is also a fundamental determinant of model robustness beyond just the architecture of its operation.

Limitations Under Harsh Visual Conditions

Despite the progress made to enhance the accuracy of autonomous vehicles in this field, the effects of environmental variability have proven to be one of the most persistent challenges for object detection systems. Studies on model performance have shown significant decreases in the accuracy of commonly used models when tested under conditions such as fog, snow, or nighttime conditions (Chauhan et al., 2024; Kumar & Muhammad, 2023). In situations where it is hard for vehicles to even see traffic signs, their chances at classifying and interpreting road marks are even lower. This threatens the potential dependability of autonomous vehicles, as limitations like these must be overcome to ensure their reliable operation. A study by Qu et al. (2023) further expands on this, noting that YOLO models are no exception, struggling to perceive occluded objects, small objects, and traffic signs under harsh weather conditions. This poses a significant problem, as YOLO models, one of the most widely used systems in the field, face clear setbacks that could pose a great risk for autonomous vehicles out on the road. The studies done by Chauhan et al., Kumar & Muhammad, and Qu et al. all recognize that this challenge persists within current literature and serves as one of the many obstacles for the reliability of this emerging technology.

Existing Approaches to Improve Model Robustness

Model Architecture and Training Approaches

In response to these limitations, current research has explored several approaches to improve model performance under harsh visual conditions. One common strategy involves modifying model architecture. Several works have proposed solutions to this by developing improved model architectures, such as versions of YOLO which optimizes its learning algorithms to better address these complex test cases (Alasmari et al., 2023). Tying back to the operation of CNNs themselves, this approach demonstrates a method where the detection process of these models is what's being focused on. Targeting model architecture appears more commonly as it's easy to compare with benchmark models and researchers can clearly define what aspects were being manipulated in the model's operation. In a study by Qiu et al. (2025), researchers developed DP YOLO, a modified version of the YOLO architecture that was specifically tailored for detecting smaller objects within the hierarchy process. They found success, with their customized model showing a significant improvement in detection accuracy when compared to baseline models. Similarly, a study by Chauhan et al. (2024) focused on using dehazing techniques and image enhancement to improve traffic detection under foggy weather. While both Qiu et al. and Chauhan et al. demonstrated measurable improvements through the manipulation of model architecture, both pertained to specific scenarios, small objects and foggy weather conditions, respectively. The underlying issue of being able to improve model robustness and generalize to a broad range of harsh conditions still remains a challenge, one that cannot be solved by scenario-specific model architecture enhancements.

The Role of Dataset Diversity

As an alternative to modifying model architecture, dataset diversity is a proposed solution which many researchers see to have promising potential. Researchers such as Lim et al. (2023) and Jadoun (2023) argue that datasets including a greater variety of lighting, weather, and environmental conditions are a key solution to more robust models. Training these systems on more diverse datasets effectively exposes them to a wider range of possible scenarios, so they are more prepared to handle complex or unusual situations. This addresses the explanation behind distribution shift and importance of training images emphasized by Mao et al. (2023). Kumar and Muhammad (2023) tested this theory, using data augmentation to merge multiple datasets together in order to create a more diverse pool of images. They found that doing so successfully improved the accuracy of YOLOv8 under adverse visual conditions. Despite these contributions, however, their study did not establish a clear measurement to define the magnitude of diversity.

Aldoski and Koren (2023) highlight this problem, that while research has been done, the field lacks quantitative measures capable of defining a clear, numerical relationship between dataset diversity and the performance of these models. With most studies relying on descriptive accounts of dataset diversity, it's hard for researchers to establish causation because of an absence in statistical evidence.

The Gap

A consistent theme emerges when analyzing the literature surrounding object detection models in autonomous vehicles: despite ongoing advancements, limitations persist regarding their performance in harsh visual conditions. Studies in the field recognize the importance of expanding environmental diversity in datasets to address this problem, but there is a lack of quantitative evidence to link the relationship between dataset diversity and model performance. This gap highlights a need to numerically and statistically quantify the correlation between these two variables. This study addresses this gap by providing a controlled experimental approach, where dataset diversity can be isolated and manipulated as the independent variable. By operationally defining the environmental diversity of training datasets used, clear measurements were taken to help determine the relationship between diversity and the accuracy of models under harsh visual conditions.

METHODS

Data Source and Image Extraction

All images used for this study were taken from the Mapillary Traffic Sign Dataset. This dataset was chosen because it is recognized as the most diverse publicly available traffic sign dataset. Its images contain a wide variety of geographic locations and environmental characteristics that will better allow for the construction of custom datasets where diversity can be manipulated (Zhao, 2026). Images from MTSD were filtered and converted using COCO format, a widely adopted layout to store the image data and annotations in a way that can easily be read by YOLO models, as seen in a previous study by Kumar and Muhammad (2023).

Environmental Metric Quantification

After the images were extracted, each image was assigned three photometric measurements in order to capture their varying features: brightness, contrast, and blur. Measurements were taken by converting the image to grayscale and then measuring the different pixel intensities as well as edge characteristics; these features represent how an image's quality would be affected by environmental factors like glare or fog (Kumar & Muhammad, 2023). Then, a harshness score was calculated based on these three other metrics in order to provide a single quantitative measure of the image's quality. It effectively measured how far away an image's quality deviates from the optimal viewing conditions of the other three measurements, using the following formula:

$$harshness = \frac{(1-contrast) + (1-blur) + |brightness-0.5|}{3}$$

Dataset Construction

Evaluation Dataset

With the measurements for each image now recorded, images were then sorted and separated into four custom datasets. The first one was an evaluation dataset, intended to represent the most challenging real world conditions which the models might encounter. This dataset, referred to as the HARSH dataset, was constructed by taking the top 4,456 images with the highest harshness scores, as seen in the red tail of Figure 1.

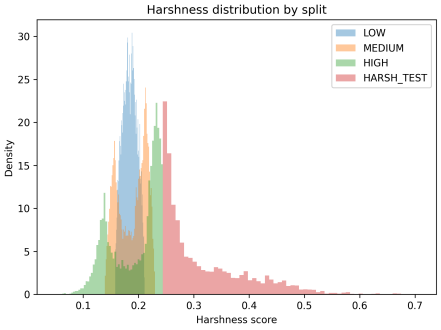

Figure 1

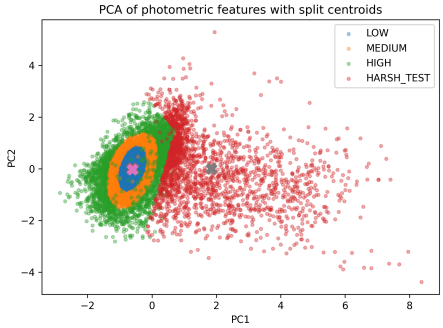

Figure 2

Training Datasets

Next, the remaining images were used to construct three training datasets with increasing levels of diversity: LOW, MEDIUM, and HIGH. Each image was represented as a four dimensional vector containing their four measurements. Then, the global centroid for this was computed, representing the "center", or average value for all these features. Figure 2 displays this located at the pink cross.

The LOW diversity dataset was composed of the 8,000 images with the smallest Euclidean distance from the centroid. In other words, its characteristics were most tightly clustered around characteristics considered "normal". This low variability represents why it as a training dataset had the least diversity within its images. Similarly, the HIGH diversity dataset contained the 8,000 images farthest away from the center, capturing the most extreme, spread out, and wide range of environmental conditions. Lastly, the remaining 8,000 images representing the intermediate distances between the LOW and HIGH images were taken to construct the MEDIUM diversity dataset. The blue, orange, and green rings in Figure 2 represent the LOW, MEDIUM, and HIGH diversity datasets, respectively. Each training dataset had the same amount of images in order to control for the variable of training size when tested on the models. In each of the training datasets, 90% (7200 images) were used for training and 10% (800 images) were used for validation. In addition, preliminary training runs were run before to monitor how many epochs were efficient. Validation mAP was compared and began to consistently stabilize after 40 epochs, supporting 50 epochs being enough to ensure convergence.

Diversity Quantification

In order to quantitatively measure the diversity of each dataset, a Vendi Score was computed (Friedman & Dieng, 2023). This was used to represent the number of distinct images within each dataset to gauge how varied it is. In other words, if a random image were to be selected, it assessed the probability of choosing another random image that would have similar qualities. A higher Vendi Score indicated greater diversity within a dataset. This measurement was chosen compared to simpler measures of variation because it accounts for the full structure of the dataset without the need for labeled categories (Friedman & Dieng, 2023).

Model Selection and Training

Three different object detection models were trained and evaluated for this study: YOLOv5s, YOLOv7-tiny, and YOLOv8n. These models were selected for being publicly available and commonly used in the field. All three belong to the YOLO family, a series of CNN models that are popular among autonomous vehicles for being accurate and able to classify information quickly under high speed situations (Qiu et al., 2025). Multiple, different models were chosen to see if the impact of diversity would prove effective across multiple types of model architecture by comparing their changes in performance with each other (Alahdal et al., 2024). v5, v7, and v8 represent different generations of the YOLO family, while their respective suffixes indicate the size of each detection model.

Each of the three models were trained on independent instances of the LOW, MEDIUM, and HIGH diversity datasets, for a total of 9 runs. Certain parameters were held constant in order to ensure diversity was the only variable being changed across these instances, including: 50 training epochs, 640x640 image size, and using pretrained COCO weights. An epoch can be defined as the amount of times each image is passed through the model during the training process. 50 epochs were chosen so that the training process would not take too long but still provide enough cycles for the models to sufficiently learn from. Pretrained weights were used in order to speed up the training process and utilize already existing training data for the backbone of the models, a common practice in training traffic sign detection models (Kumar & Muhammad, 2023).

All nine instances of the models trained were evaluated using the HARSH dataset test. The accuracy of these models was assessed using four standard object detection metrics: precision, recall, mAP@50, and F1 score. Precision measures the proportion of the model's detections that were correct, while recall measures the proportion of actual traffic signs in the image that the model successfully detected. F1 score combines both into a single balanced metric, accounting for the tradeoff between the two. mAP@50 aggregates precision across all object classes at a 50% overlap threshold, providing a comprehensive single measure of overall detection accuracy (Qiu et al., 2025). These metrics were chosen because of their prior use in other studies and being established benchmarks for measurements of accuracy in the field of object detection models (Qiu et al., 2025; Kozhamkulova et al., 2024).

Statistical Analysis

Both descriptive and inferential statistics were used to analyze the data. First, inferential statistics including an ANOVA test (Wang et al., 2025) was used to confirm whether the difference in diversity levels between training datasets was manipulated significantly as an independent variable. Similarly, effect size was reported using eta-squared to quantify the difference in proportions of diversity. Descriptive statistics including the standard deviation of harshness within image datasets was also recorded as another way to back up the diversity represented by Vendi Score.

As for the evaluation of the models, percent change in accuracy metrics were recorded for easy interpretation of whether or not diversity had significant change on the model's performance. In addition, measurements such as the absolute range for measurements were taken to signify which model experienced the greatest impact due to changing diversity. Specifically, mAP@50 was used as the baseline measurement to compare how accurate the models performed under the HARSH test dataset.

RESULTS AND ANALYSIS

Dataset Diversity Validation

Prior to evaluating each of the models, tests were run to ensure that the manipulation of diversity produced meaningful differences between the datasets. Figure 3 summarizes the key metrics across all four datasets, including mean harshness, standard deviation of harshness, and Vendi Score for each dataset.

Figure 3

Dataset	Images	Avg Harshness	SD Harshness	Vendi Score
LOW	8000	0.1840	0.0127	2.4223
MEDIUM	8000	0.1879	0.0259	2.4674
HIGH	8000	0.1922	0.0439	2.8476
HARSH	4456	0.3181	0.0791	3.1003

As seen in the table, both the standard deviation of harshness and Vendi Score increased continuously across the increasing levels of diversity for the datasets. This demonstrates that the procedure for separating and organizing the images using their 4 metrics was successful in constructing datasets with increasing levels of diversity. In addition, the results demonstrated that the HARSH evaluation dataset had the highest values for all measures, which aligned with the goal of it being a representation of the most

visually challenging images. A one way ANOVA was conducted for the harshness scores of images between datasets, and it confirmed that the change in diversity was statistically significant. The p-value to have occurred by chance was nearly zero, and the eta-squared value was 0.5613, indicating that 56% of the variance in harshness across all images could be attributed to the manipulation of the datasets.

YOLOv8n Results

Figure 4

Dataset	Precision	Recall	F1 score	mAP50
LOW	0.7580	0.4770	0.5855	0.5400
MEDIUM	0.7476	0.4800	0.5847	0.5418
HIGH	0.7433	0.4851	0.5871	0.5427

The table above depicts YOLOv8n's performance when evaluated using the HARSH test dataset. It saw a slight continuous increase in mAP@50 as diversity increased, but only by an absolute difference of 0.0027 from LOW to HIGH, approximately 0.5%. Recall increased slightly, while precision declined a marginal amount. F1 score stayed relatively the same though it did vary a bit with no clear trend. The effect size for all these changes were negligible, indicating that the performance was not significantly affected by increasing the diversity of the training datasets used.

YOLOv5s Results

Figure 5

Dataset	Precision	Recall	F1 score	mAP50
LOW	0.775	0.535	0.633	0.607
MEDIUM	0.774	0.528	0.628	0.601
HIGH	0.767	0.528	0.625	0.600

Next, Figure 5 depicts the performance of YOLOv5s when evaluated under the HARSH test dataset. For this model, a slight negative trend was seen across all of the metrics, with mAP@50 declining a total of 0.007 from LOW to HIGH diversity, about 1.2%. Precision, Recall, and F1 score followed the same

pattern, but only by marginal differences. While it's worth noting that YOLOv5s displayed the highest accuracy compared to the performance of other models, the negligible impact of changing diversity level served to refute support of the hypothesis.

YOLOv7-tiny Results

Figure 6

Dataset	Precision	Recall	F1 score	mAP50
LOW	0.7420	0.5240	0.6142	0.5800
MEDIUM	0.7630	0.5310	0.6262	0.5960
HIGH	0.7360	0.5320	0.6176	0.5880

Lastly, Figure 6 depicts the performance of YOLOv7-tiny when evaluated under the HARSH test dataset. The model experienced no clear trend as the diversity of the training dataset increased. mAP@50 increased about 0.016 from LOW to MEDIUM, but then decreased back 0.008 from MEDIUM to HIGH. The other metrics followed a similar pattern, with the MEDIUM training dataset producing the highest precision, recall, and F1 score as well. This model saw the largest absolute difference in mAP@50, about 0.016 between LOW and MEDIUM, but still negligibly small in magnitude. This result doesn't support any positive or negative trend between dataset diversity and the accuracy of the model when evaluated under the HARSH dataset.

Cross Model Comparison

Across all 3 models, no significant relationship was observed between the diversity levels of datasets and the model performance under the HARSH evaluation dataset.

Figure 7

Model	LOW	MEDIUM	HIGH	Range	Direction
YOLOv8n	0.5400	0.5418	0.5427	0.0027	Weak positive
YOLOv5s	0.6070	0.6010	0.6000	0.0070	Weak negative
YOLOv7-tiny	0.5800	0.5960	0.5880	0.0160	Varying trend

The direction of the observed trends was also inconsistent across the models. YOLOv8n showed a slight positive trend, YOLOv5s slightly negative, and YOLOv7-tiny was varied, peaking at MEDIUM diversity. The absence of any clear relationship across all 3 types of architectures refutes the hypothesis that increasing diversity of the training datasets had any significant impact on the accuracy of these models when tested under extreme harsh visual conditions. The largest observed range in difference for mAP@50 was 0.016, only a 2.8% increase in performance during the tests for the YOLOv5s model. This effect size is not significant and small relative to differences reported in previous work manipulating model architecture (Qiu et al., 2025). Consequently, the results do not provide convincing evidence in support of the hypothesis, instead demonstrating that changing diversity of training datasets does not produce meaningful improvements for model robustness under visual degradation.

DISCUSSION

Overall, the study found that manipulating the diversity of training datasets did not display any relationship with the performance of the models when evaluated under harsh visual conditions. This held true for all 3 models: YOLOv8n, YOLOv5s, and YOLOv7-tiny, as there was no clear trend in any of their metrics for precision, recall, mAP@50, or F1 score. These findings refuted the hypothesis that increased diversity would lead to increased accuracy.

Comparison With Previous Studies

The findings of these results partially refute the claims of Kumar and Muhammad (2023), who largely influenced the hypothesis. They found that merging diverse datasets improved the robustness of the YOLOv8 model, implying that increased diversity did increase accuracy. However, it is worth noting that the methods were different, as they used data augmentation and merged multiple separate datasets, while this study used one dataset and manipulated data by sorting them into increasing Vendi scores. In addition, it possibly provides further insight into the exact technicalities of their findings. A possible cause of the limited effect due to manipulating diversity was that the dataset sizes were not extremely large, as used in Kumar and Muhammad's
(2023) study. This would be in agreement with a study by Rajput (2023), who mentions that both too large or too small sample sizes can have negative effects on the training processes of models, including problems like overfitting. This suggests an altered proposal to the findings of Kumar and Muhammad (2023): that simply increasing the diversity of training datasets without accounting for other variables will not automatically increase accuracy, but it's also important to control other factors like dataset size. In addition, the findings support the common theme in the field of many large studies choosing to enhance model architecture, providing much larger results and significant improvements in accuracy compared to this study (Karbouj et al., 2026).

Implications

While the hypothesis was not supported, this study still holds significant implications for the field. For one, it suggests that those seeking to improve the accuracy of models under harsh visual conditions should not aim to do so by only increasing environmental diversity alone. It's important to also control for other factors such as dataset size or even the use of pretrained weights. The findings of this study suggest that a larger dataset size is necessary in order for the effects of image diversity to fully manifest within the results of models' performances. In addition, it validates that resources may be better directed to the manipulation of model architecture as a means of enhancing the performance of these models. In other words, research into the actual detection process of these models, such as Qiu et. al's (2025) modification of DP YOLO, yields more significant improvements which should be further researched for the field.

Still, the methodology used for this study is generalizable for a display of quantitatively measuring diversity of datasets in the broader field. It shows a way to systematically break up a dataset based on assigned environmental metrics and sort them based on their characteristics, a procedure reproducible and applicable to any form of object detection involving datasets. Furthermore, the study also makes significant contributions as an example of the scientific value in reporting null results. In an article by Karl et al. (2024), they explain that many large publications in machine learning only display positive, "impressive" results which claim to be current field standards. This calls the need for the publication of more "negative" results, or authentic results that add meaningful contributions to the field by showing what does or doesn't work, especially in this field. The current study provides an exact means to this as the manipulated diversity shown wasn't a clear cut solution towards improving accuracy, but revealed other possible variables, like dataset size that needed to also be considered and how this could be improved for future studies. Such results reinforce the iterative process that comes with many fields in STEM, where research is constantly building off of new findings to help experts better understand the functioning and flaws of these machine learning models.

Consequently, this study offers useful insight to researchers in the field of computer vision, autonomous vehicle development, and machine learning. Its results provide useful findings for those that develop contemporary traffic sign recognition systems, by providing a deeper understanding on how diversity can affect the performance of these models under robust conditions. Organizations that work on creating and standardizing datasets can also take note of the use of Vendi score to measure diversity and how more robust training sets can be optimized for the future.

Limitations

Connecting to this, limitations of this study were certainly present. As mentioned early, the relatively small dataset sizes (n=8000) may have not been large enough to produce significant distributional differences of the training process for each of the models. It's possible that the effect of diversity could manifest in the study of dataset magnitude was much larger, yet this study was limited by access to

storage and memory of images being able to be stored on the researcher's computer. Another possible limitation was the use of pretrained COCO weights. This is often common practice used to kickstart the training process for models in the field, with many autonomous vehicles on the roads utilizing this method; however, it's possible that the imposed influence of this initialization may have fed too much training into the models that practically would already override any of the later effects which diversity had in the training process. A study conducted where models are trained from scratch would yield interesting results but certainly require more time and computational resources that were limited in this study. The study could also be more robust if multiple trials/trainings were included for each model rather than just one training and evaluation for each model with each dataset. This would allow for the calculation of more inferential statistics that could better capture the true relationship and trends observed between the two variables. In addition, averages could be taken from the trials to ensure a more concrete result not swayed by the possibly randomness caused from a single training run.

Conclusion

This study analyzed the relationship between training dataset diversity and the accuracy of object detection models when tested under harsh visual conditions. Numerical metrics were used to measure both variables in order to fill a gap in the current field, a lack of quantitative correlation between the two. The hypothesis was that greater dataset diversity would exhibit a positive relationship with the tested models' performance. Three models were tested: YOLOv8n, YOLOv5s, and YOLOv7-tiny. Each model was constructed on three custom datasets of LOW, MEDIUM, and HIGH level diversity, measured by dividing up the harsh environmental attributes of images and recording a photometric Vendi score. All images were taken from the Mapillary Traffic Sign Dataset, and each training instance was evaluated on a separate HARSH dataset to compare their performances using precision, recall, F1 score, and mAP@50. The results of the study found no consistent or statistically significant effect between the two variables. While statistical tests, including a one way ANOVA, were run to confirm that the manipulation of the diversity was large, its effect when each of the models evaluated showed no clear correlation. This lack in pattern appeared across all 3 model architectures, suggesting that changing the diversity had little impact when multiple types of models were tested under harsh test cases. Such findings refuted the hypothesis, failing to provide convincing evidence that under the conditions of this experiment did increasing the environmental photometric diversity of training datasets improve model's performance in harsh visual conditions.

However, the study's findings still carry significant contributions towards the field, one with the demonstration of the use of a Vendi score and harshness scores to manipulate diversity. While no relationship was proven, the methods of this study quantifying diversity can be used in future studies who further wish to explore the implications of numerically measuring dataset diversity and its effects. In addition, while new techniques to improve model robustness were not proven, that does not mean such findings should not be contributed to the field. Rather, it demonstrates what did not work, revealing limitations to be addressed and further expanded on for future studies in the field.

Future Directions

As mentioned earlier, one limitation of the study was the dataset scales. Each contained about 8000, and at this sample size diversity was confirmed to not have a significant effect on increasing model robustness. For future studies, this is a factor certainly worth investigating. Increasing the dataset scale to a much larger size–for example 50,000 or more per training dataset–could reveal different results and possibly uncover promising implications of diversity not discovered at a small scale. In addition, future experiments should incorporate multiple trials for each instance of training. This would entail training each model on each training dataset multiple times, and holding an evaluation for each of those trials. Such rigor could confirm the trend being observed is consistently occurring, and more inferential statistical analysis could be applied to back this up. Given more resources, future studies could also examine the effect of different types of model architectures beyond just YOLO models. This would extend to not only other CNNs, but even two stage detectors like Faster R-CNN, which also are prominently used in autonomous vehicles but have a completely different learning process. It would be interesting to observe whether these vastly different architectures, compared to the differences within one family of models, displays a greater contrast in results. By addressing the limitations noted within this study and applying improvements to the methods of this experiment, significant findings can be expanded upon for this field to enhance the reliability and accuracy of traffic sign detection models.

REFERENCES

Alahdal, N. M., Abukhodair, F., Meftah, L. H., & Cherif, A. (2024). Real-time object detection in autonomous vehicles with YOLO. *Procedia Computer Science, 246,* 2792–2801. https://doi.org/10.1016/j.procs.2024.09.392

Alasmari, N., Alohali, M. A., Khalid, M., Almalki, N., Motwakel, A., Alsaid, M. I., Osman, A. E., & Alneil, A. A. (2023). Improved metaheuristics with deep learning based object detector for intelligent control in autonomous vehicles. *Computers and Electrical Engineering, 108,* 108718. https://doi.org/10.1016/j.compeleceng.2023.108718

Aldoski, Z. N., & Koren, C. (2023). Impact of traffic sign diversity on autonomous vehicles. Periodica Polytechnica Transportation Engineering, 51(4), 338–350. https://doi.org/10.3311/pptr.21484

Chauhan, S., Ghule, P., Deshmukh, A., & Bhagwat, O. (2024). Traffic sign detection under foggy conditions using machine learning. *International Journal of Advanced Research in Science, Communication and Technology,* 370–375. https://doi.org/10.48175/ijarsct-22354

Contreras, M., Jain, A., Bhatt, N. P., Banerjee, A., & Hashemi, E. (2024). A survey on 3D object detection in real time for autonomous driving. *Frontiers in Robotics and AI, 11*. https://doi.org/10.3389/frobt.2024.1212070

Friedman, D., & Dieng, A. B. (2023). The Vendi Score: A Diversity Evaluation Metric for Machine Learning. Alphaxiv.org; arXiv. https://www.alphaxiv.org/overview/2210.02410v2

Jadoun, P. (2023). Traffic sign detection and recognition system for autonomous vehicle.

International Journal for Research in Applied Science and Engineering Technology, 11(11), 1013–1017. https://doi.org/10.22214/ijraset.2023.56638

Karbouj, B., Altenbuchner, A. M., & Krueger, J. (2026). Deep Learning-Based Object Detection for Autonomous Vehicles: A Comparative Study of One-Stage and Two-Stage Detectors on Basic Traffic Objects. ArXiv.org. https://arxiv.org/abs/2602.00385

Karl, F., Kemeter, Lukas Malte, Dax, G., & Sierak, P. (2024). Position: Embracing Negative Results in Machine Learning. ArXiv.org. https://arxiv.org/abs/2406.03980

Kozhamkulova, Z., Bidakhmet, Z., Vorogushina, M., Tashenova, Z., Tussupova, B., Nurlybaeva, E., & Kambarov, D. (2024). Development of Deep Learning Models for Traffic Sign Recognition in Autonomous Vehicles. International Journal of Advanced Computer Science and Applications, 15(5). https://doi.org/10.14569/ijacsa.2024.0150593

Kumar, D., & Muhammad, N. (2023). Object detection in adverse weather for autonomous driving through data merging and YOLOv8. *Sensors, 23*(20), 8471. https://doi.org/10.3390/s23208471

Landge, A., Marotkar, S., Shelar, R., Kolte, K., & Babar, V. (2025). *A review on object detection for autonomous vehicles*. IJRASET. https://www.ijraset.com/research-paper/object-detection-for-autonomous-vehicles

Lim, X. R., Lee, C. P., Lim, K. M., Ong, T. S., Alqahtani, A., & Ali, M. (2023). Recent advances in traffic sign recognition: Approaches and datasets. *Sensors, 23*(10), 4674. https://doi.org/10.3390/s23104674

Mao, X., Chen, Y., Zhu, Y., Chen, D., Su, H., Zhang, R., & Xue, H. (2023). COCO-O: A Benchmark for Object Detectors under Natural Distribution Shifts. ArXiv.org. https://arxiv.org/abs/2307.12730

Qiu, J., Zhang, W., Xu, S., & Zhou, H. (2025). DP-YOLO: A lightweight traffic sign detection model for small object detection. *Digital Signal Processing,* 105311. https://doi.org/10.1016/j.dsp.2025.105311

Qu, S., Yang, X., Zhou, H., & Xie, Y. (2023). Improved YOLOv5-based for small traffic sign detection under complex weather. Scientific Reports, 13(1), 16219. https://doi.org/10.1038/s41598-023-42753-3

Rajput, D., Wang, W.-J., & Chen, C.-C. (2023). Evaluation of a decided sample size in machine learning applications. BMC Bioinformatics, 24(1). https://doi.org/10.1186/s12859-023-05156-9

Shen, J., Zhang, Z., Luo, J., & Zhang, X. (2023). YOLOv5-TS: Detecting traffic signs in real-time. *Frontiers in Physics, 11*. https://doi.org/10.3389/fphy.2023.1297828

Wang, C., Zheng, B., & Li, C. (2025). Efficient traffic sign recognition using YOLO for intelligent transport systems. Scientific Reports, 15(1). https://doi.org/10.1038/s41598-025-98111-y

Youssouf, N. (2022). Traffic sign classification using CNN and detection using faster-RCNN and YOLOV4. Heliyon, 8(12), e11792. https://doi.org/10.1016/j.heliyon.2022.e11792

Zhao, G., Qi, W., Zhang, K., Zhang, C., Gong, Z., Bi, Z., Chen, K., Ma, B., Liu, M., & Ma, J. (2026). Traffic Sign Recognition in Autonomous Driving: Dataset, Benchmark, and Field Experiment. ArXiv.org. https://arxiv.org/abs/2603.23034