Proceedings of CONF-SEML 2026 Symposium: Computational Analysis and Modeling in Complex Intelligent Systems
DOI: 10.54254/2755-2721/2026.GU33437

Assessing the Robustness of Traffic Sign Classification under Image Degradation

Ziying Yang

School of Computer Science, University of Nottingham, Nottingham, United Kingdom
scyzy12@nottingham.edu.cn

Abstract. The robustness of a traffic sign classification model was evaluated under degraded test conditions in a controlled Edge Impulse workflow. The experiments were conducted on a balanced subset of the German Traffic Sign Recognition Benchmark dataset. Eight traffic sign classes were selected, and each class contained 700 training images and 300 test images. A MobileNetV2-based transfer learning model was trained only on clean images, and image degradation was added only to the test set. Gaussian noise, salt-and-pepper noise, and Gaussian blur were applied to the test images. Each type had low, medium, and high levels. Evaluation was based on overall accuracy, weighted metrics, and confusion matrices. On the clean test set, baseline accuracy reached 98.36%, but performance decreased clearly under all degraded conditions. Of the three degradation types, Gaussian noise had the largest negative effect, and Gaussian blur had the smallest overall effect. The confusion matrix results also showed more misclassification under severe degradation, especially between visually similar classes. These results show that strong performance on clean benchmark data does not always mean the model will also work well in more realistic traffic sign recognition conditions.

Keywords: Traffic sign classification, model robustness, image degradation, Edge Impulse

1. Introduction

In intelligent transportation systems, traffic sign classification matters because it helps driver assistance systems and autonomous vehicles to recognize road signs correctly [1]; reliable recognition is needed not only for a correct reading of traffic rules but also for road safety [2]. Real-world traffic sign recognition however remains challenging for two reasons: one is that traffic signs may show up in complex scenes and under changing lighting conditions, and another is that image quality can also get affected by other environmental factors [1,2].

Deep learning has turned into a common approach for image classification and is now widely used in traffic sign recognition [1]. Convolutional neural networks in particular fit this task well because they can learn visual information straight from traffic sign images [3]; this helps the model to tell signs apart using features like shape, color, and symbols. Earlier studies have shown strong performance under standard testing conditions on benchmark datasets, including the German Traffic Sign Recognition Benchmark (GTSRB) [3]; because of that, benchmark datasets are often used to

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2026.GU33437

compare different traffic sign recognition models. Still, a strong result on benchmark datasets does not always mean that a model is robust in real-world environments [1,2].

Despite these promising results, most evaluations are still done under standard or controlled conditions [1,4]; such results do not necessarily reflect how models behave when test images get degraded in more practical scenarios. In real applications, test images may suffer from various sources of degradation during acquisition and deployment; one point to note is that the effect of image noise on classification performance has not always been examined in a simple and systematic way. Understanding how these factors affect classification performance is therefore important for deciding whether a model can stay reliable outside ideal benchmark settings; this makes a controlled robustness evaluation especially valuable in a deployment-oriented workflow such as Edge Impulse.

This work focuses on the robustness of a traffic sign classification model under degraded test conditions. To keep the comparison controlled, the model was trained only on clean images and was not altered during testing; only the test set was modified by applying different noise types. The evaluation relied on overall accuracy together with weighted recall, weighted precision, weighted F1-score, and confusion matrix analysis.

2. Method

2.1. Dataset preparation

The experiments focused on multi-class traffic sign classification and were based on the GTSRB dataset from Kaggle [5]; this dataset contains more than 50,000 images from over 40 traffic sign classes and is widely used as a benchmark in traffic sign classification research. It was suitable for this experiment for two reasons: one is that the selected classes are visually clear, and another is that this clarity made the effect of noise on classification performance easier to observe.

Eight typical classes were selected from the original 42 classes in GTSRB, namely General caution, Keep right, No entry, Right-turn ahead, Road work, Roundabout mandatory, the 50 km/h speed limit sign, and the 60 km/h speed limit sign. These classes cover different sign types including speed limit signs, warning signs, prohibition signs, and mandatory signs; the original image sizes were not the same across these classes.

To ensure a fair comparison across classes, the experiments used a balanced dataset; for each selected class, 1,000 images were prepared with 700 images for training and 300 images for testing. This fixed split kept class sizes consistent and reduced the influence of class imbalance during evaluation; using the same number of training and test images in each class also made comparisons across different noise conditions more direct.

Before training, all images were resized to a fixed input size. Color plays an important role in traffic sign recognition, so the images were stored in RGB format to keep the color information; after that, the images were organized by class labels and prepared for model training and testing.

The clean test set was kept unchanged and used as the baseline test set; then corrupted versions of the same test images were created. This design kept the training condition fixed, which allowed the effect of image degradation on model robustness to be examined more directly. Three degradation types were applied to the test images: Gaussian noise, salt-and-pepper noise, and Gaussian blur; each type had three severity levels which were low, medium, and high. The detailed parameter settings for all degraded test conditions are shown in Table 1, and example images of the clean test set together with the degraded test conditions are shown in Figure 1. For Gaussian noise, the low, medium, and high levels used standard deviation values of 10, 20, and 30 respectively; for salt-and-pepper noise, the three levels used corrupted pixel ratios of 1%, 3%, and 5%; for Gaussian blur, the

DOI: 10.54254/2755-2721/2026.GU33437

three levels used kernel sizes of 3×3, 5×5, and 7×7. These degraded test sets were then used to assess how robust the trained model remained under different image quality conditions.

Table 1. Parameter settings for degraded test conditions

Degradation type	Low	Medium	High
Gaussian noise	σ = 10	σ = 20	σ = 30
Salt-and-pepper noise	1% corrupted pixels	3% corrupted pixels	5% corrupted pixels
Gaussian blur	3 × 3 kernel	5 × 5 kernel	7 × 7 kernel

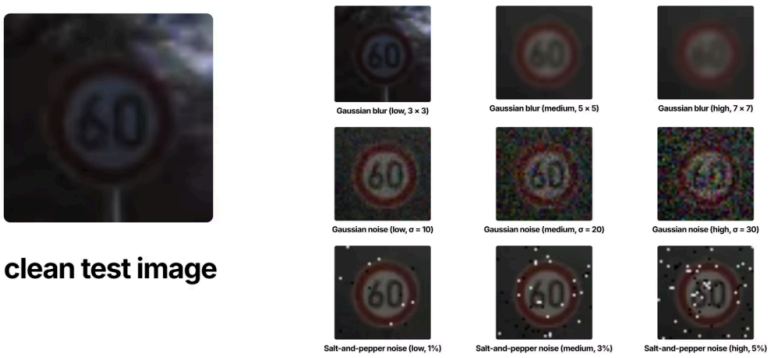

Figure 1. Example clean (left) and degraded (right) test images (picture credit: original)

2.2. Edge impulse-based CNN model

Convolutional neural networks are well suited to image classification because they can learn useful visual patterns directly from images. By using convolutional layers, they can extract local features such as edges, textures, and shapes automatically [4]. This makes them suitable for traffic sign classification because traffic signs are distinguished by important visual features such as color, shape, and internal pictogram content [6,7].

Edge Impulse was used for model training and evaluation [8]. The labeled image folders were uploaded to the platform. Then a transfer learning image classification project based on MobileNetV2 was created with an input size of 160×160 pixels [9]. The platform provided a clear and structured pipeline with consistent settings, so it supported controlled comparisons in this course-based experiment. Training was then carried out on the prepared dataset.

The same trained model and the same evaluation pipeline were used for all test conditions. The experiments used a MobileNetV2-based transfer learning model with 20 training cycles and a learning rate of 0.0005. Model training was carried out on the CPU. The other training and model settings followed the default settings in Edge Impulse.

The evaluation metrics included overall accuracy, weighted F1-score, weighted precision, and weighted recall. Confusion matrix analysis was also carried out to inspect class-level prediction outcomes and reveal misclassification patterns. The final analysis covered both baseline results from the original test images and robustness results from degraded test images under different noise conditions.

DOI: 10.54254/2755-2721/2026.GU33437

3. Results and discussion

3.1. Baseline results on the clean test set

The clean test set was used as the baseline condition. On the clean test set, overall accuracy reached 98.36%. Weighted precision, recall, and F1-score each reached 0.99. The class-level prediction results are shown in Figure 2. Under the clean test condition, most traffic sign classes were classified correctly, and only a few misclassifications appeared. The best class-level results were for No entry and Roundabout mandatory, and both reached 100% classification accuracy. Small errors still appeared in a few visually similar classes, especially the two speed limit categories. These results from the clean test set were used as the reference point for all later comparisons under degraded conditions.

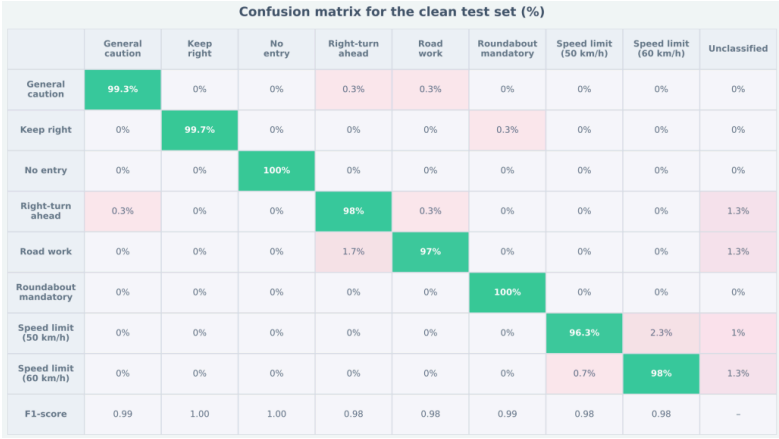

Figure 2. Confusion matrix for the clean test set (picture credit: original)

3.2. Comparison across different noise types

Different noise types affected the model in different ways at the overall level. As shown in Table 2, when performance was averaged across the three severity levels, Gaussian noise gave the weakest overall results, with the lowest average accuracy of 45.43%. This shows that the model was least robust under Gaussian noise. Salt-and-pepper noise gave middle-level performance, with an average accuracy of 56.58%. Gaussian blur gave the best average results among the three degraded conditions, with an average accuracy of 78.26%. But this was still clearly lower than the clean baseline. The same trend was also observed for weighted precision, weighted F1-score, and recall.

One possible reason is that Gaussian noise introduced random perturbations across the whole image, which hindered the extraction of discriminative visual features and made the classes harder to separate [10]. Salt-and-pepper noise also damaged the image, but its overall effect was smaller than Gaussian noise in this study. Gaussian blur mainly weakened edge information and other fine visual details [11]. These details are important for classification, but the model still performed better under blur than under the other two degradation types.

DOI: 10.54254/2755-2721/2026.GU33437

Table 2. Average classification performance across different noise types

Degradation type	Accuracy	Weighted precision	Weighted recall	Weighted F1-score
Gaussian noise	45.43%	0.69	0.48	0.49
Salt-and-pepper noise	56.58%	0.74	0.59	0.58
Gaussian blur	78.26%	0.84	0.80	0.81

3.3. Comparison across noise severity levels

Model performance became worse as noise severity increased for all three noise types. As shown in Figure 3 and Table 3, for Gaussian noise, accuracy dropped from 67.46% at the low level to 41.33% at the medium level and 27.50% at the high level. For salt-and-pepper noise, accuracy also dropped clearly, from 77.79% to 52.38% and then to 39.58%. Gaussian blur showed the same trend, with accuracy dropping from 89.66% at the low level to 80.43% at the medium level and 64.70% at the high level. The other weighted metrics also showed the same pattern.

This pattern shows a clear severity effect across all three degradation types. As the corruption level increased from low to high, classification performance dropped in every case. The biggest drop happened under Gaussian noise. This shows that the model was very sensitive to stronger random pixel-level changes. Salt-and-pepper noise also caused a clear drop, but the change was a little smaller. But Gaussian blur showed the slowest drop, suggesting that the model may still have benefited from some overall shape cues [12]. Overall, the results show that stronger degradation consistently made it harder for the model to tell traffic sign classes apart.

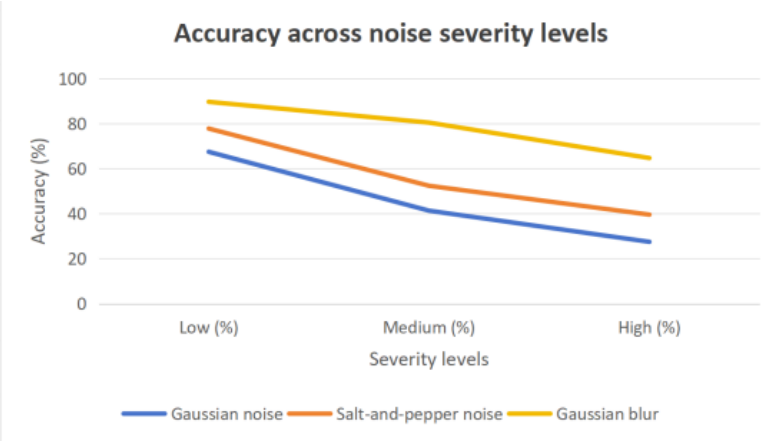

Figure 3. Accuracy across noise severity levels (picture credit: original)

Table 3. Accuracy under different degradation types and severity levels

Degradation type	Low (%)	Medium (%)	High (%)
Gaussian noise	67.46	41.33	27.50
Salt-and-pepper noise	77.79	52.38	39.58
Gaussian blur	89.66	80.43	64.70

DOI: 10.54254/2755-2721/2026.GU33437

3.4. Confusion matrix analysis under severe degradation

The confusion matrix in Figure 4 shows that severe degradation caused many more class-level misclassifications. In the clean test set, most predictions were concentrated on the diagonal, which indicates that the classes were generally well separated. However, under high Gaussian noise, the diagonal values dropped clearly for several classes and off-diagonal errors increased.

Some classes were more affected than others. Under this condition, No entry remained the most stable class, with an accuracy of 93.7%. Keep right and Speed limit (60 km/h) also remained relatively robust, while Road work and Roundabout mandatory showed much lower accuracies.

Several clear misclassification patterns also appeared. Roundabout mandatory was predicted as Keep right in 28.2% of cases, while the 50 km/h speed limit sign was predicted as the 60 km/h speed limit sign in 20.0% of cases. In addition, Road work and Right-turn ahead were frequently assigned to the Unclassified category.

These results suggest that severe Gaussian noise did not lead to fully random errors. Instead, the model may still have relied on some remaining global visual cues, but these were often insufficient for classes that also depended on local or inner sign features [13].

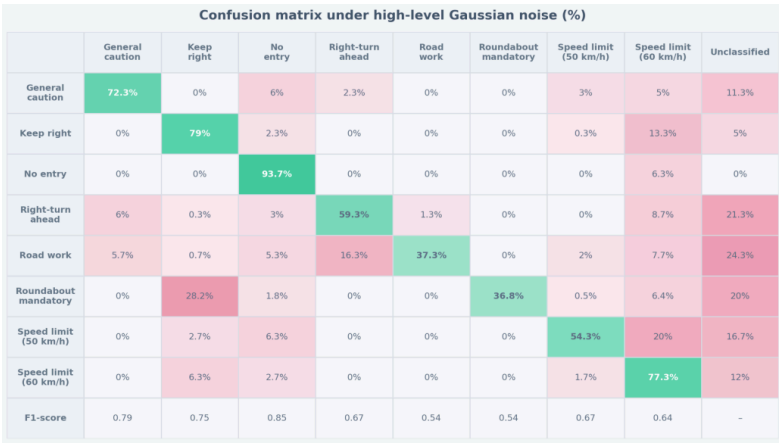

Figure 4. Confusion matrix under high-level Gaussian noise (%) (picture credit: original)

4. Conclusion

In conclusion, this study examined the robustness of a traffic sign classification model under different degraded test conditions in Edge Impulse. The model was trained only on clean images. Only the test set was modified by applying three degradation types: Gaussian noise, salt-and-pepper noise, and Gaussian blur. This design allowed the effect of image degradation on classification performance to be examined more directly. The results showed that the model had very strong baseline performance on the clean test set, with an accuracy of 98.36%, but its performance dropped clearly under all degraded conditions. Of the three degradation types, Gaussian noise caused the biggest drop, and Gaussian blur caused the smallest drop overall. The confusion matrix also showed that degradation increased misclassification between classes, especially for classes that depended more on small visual details. This means that strong performance on a clean benchmark does not always mean strong robustness in real-world traffic sign recognition conditions. One limitation of this study is that it only used eight classes and one MobileNetV2-based model. In future work, more traffic sign classes, more model types, and more real-world degradation conditions can be added to give a wider test of model robustness.

DOI: 10.54254/2755-2721/2026.GU33437

References

[1] Lim, X.R., et al. (2023) Recent advances in traffic sign recognition: approaches and datasets. Sensors, 23(10), 4674.
[2] Zhao, R., et al. (2024) Enhancing autonomous driving safety: A robust traffic sign detection and recognition model TSD-YOLO. Signal Processing, 225, 109619.
[3] Megalingam, R.K., et al. (2023) Indian traffic sign detection and recognition using deep learning. International Journal of Transportation Science and Technology, 12(3), 683-699.
[4] Triki, N., Karray, M. and Ksantini, M. (2023) A real-time traffic sign recognition method using a new attention-based deep convolutional neural network for smart vehicles. Applied Sciences, 13(8), 4793.
[5] Stallkamp, J., et al. (2011) The German traffic sign recognition benchmark: a multi-class classification competition. Proceedings of the 2011 International Joint Conference on Neural Networks, 1453-1460.
[6] Madani, A. and Yusof, R. (2018) Traffic sign recognition based on color, shape, and pictogram classification using support vector machines. Neural Computing and Applications, 30(9), 2807-2817.
[7] Vennelakanti, A., et al. (2019) Traffic sign detection and recognition using a CNN ensemble. Proceedings of the 2019 IEEE International Conference on Consumer Electronics (ICCE), 1-5.
[8] Diab, M.S. and Rodriguez-Villegas, E. (2022) Performance evaluation of embedded image classification models using Edge Impulse for application on medical images. Proceedings of the 2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC), 2464-2467.
[9] Sandler, M., et al. (2018) MobileNetV2: Inverted residuals and linear bottlenecks. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 4510-4520.
[10] da Costa, G.B.P., et al. (2016) An empirical study on the effects of different types of noise in image classification tasks. arXiv preprint arXiv: 1609.02781.
[11] Shao, F., et al. (2018) Real-time traffic sign detection and recognition method based on simplified Gabor wavelets and CNNs. Sensors, 18(10), 3192.
[12] Wali, S.B., et al. (2015) An automatic traffic sign detection and recognition system based on colour segmentation, shape matching, and SVM. Mathematical Problems in Engineering, 2015(1), 250461.
[13] Yuan, X., et al. (2014) Robust traffic sign recognition based on color global and local oriented edge magnitude patterns. IEEE Transactions on Intelligent Transportation Systems, 15(4), 1466-1477.