Explainable and Uncertainty-Aware Deep Learning for Image Classification under Noisy and Augmented Conditions

ASHWANI KUMAR AGGARWAL
Electrical and Instrumentation Engineering Department
Sant Longowal Institute of Engineering and Technology
SLIET, Longowal, 148106, Punjab
INDIA

Abstract: - The deep learning models used for image classification in various applications of computer vision such as intelligent transportation systems, precision agriculture, medical imaging, and remote sensing, etc. perform well when the dataset used for training the models is clean meaning the images in the dataset are free from noise, distortion, motion blur, occlusions, and augmented data. However, in the presence of noisy data or augmented data, the performance of many deep learning models degrades significantly, resulting in false image classification. In this paper, the performance of various deep learning models in the presence of noise and data augmentation is evaluated on benchmark datasets. Monte Carlo dropout is used for uncertainty quantification and Grad-CAM is used for the visual explainability. The performance of the models is evaluated using performance metrics accuracy and uncertainty-Robustness Index (URI). Experimental results show that the method improves the robustness of the models in the presence of noisy data.

Key-Words: - Image classification, Monte Carlo dropout, Visual explainability, Data augmentation, Uncertainty quantification.

Received: March 25, 2025. Revised: July 4, 2025. Accepted: August 7, 2025. Published: October 22, 2025.

1 Introduction

The deep learning models are widely used in image classification, segmentation tasks for applications in remote sensing, medical imaging, intelligent transportation systems, and mobile robotics. The commonly used model for such applications is convolutional neural network (CNN). The model works well in many cases when the data used for training the model is clean meaning the data is free from sensor noise or motion blur. Also, data augmentation may cause degradation of model performance in many cases. The computer vision models need to be robust in real time situations where the images are captured using sensors which have their inherent limitations and cause sensor noise.

The conventional deep learning models are treated as black boxes in which the reasons for choosing a particular class or a particular prediction made by the deep learning model is not explained. In that case, it becomes less trustworthy, and convincing to accept the output given by the model. It is especially true in applications such as medical imaging for medical image classification and in autonomous driving for image segmentation tasks. To address these issues of using deep learning models in noisy and augmented data, uncertainty estimation and explainable AI methods are discussed in the paper. The uncertainty quantification is done using Monte Carlo dropout and visual explanation of the model is done using gradient weighted class activation mapping (Grad-CAM). The performance of the model for benchmark datasets is evaluated using performance metrics such as uncertainty-robustness index (URI).

In this paper, the comparison of performance of deep learning models in the presence of noisy and augmented data is done using benchmark datasets. Monte Carlo dropout to quantify the prediction uncertainty and Grad-CAM for visual explanation of the models is discussed. The performance of the models is evaluated using uncertainty-robustness index (URI).

2 Related Work

A framework to improve the reliability of deep learning models for image classification using

uncertainty quantification and explainable AI is used in [1]. The framework focuses on classification models without discussion on its applicability in different scenarios. Monte Carlo dropout method and Grad-CAM are used in the paper for improving the robustness of the models in the presence of noisy data. The trade-off between accuracy and estimation of prediction uncertainty is discussed in [2]. The method evaluates the performance of various prediction uncertainty estimation methods used in deep learning models for image classification. The method is evaluated to CNN and limited number of uncertainty estimation models. The quantification of uncertainty in deep learning models is done using gradient information in [3]. The magnitude and the direction of the gradients is calculated for the quantification. The method is used for image classification tasks. A review of quantification methods for prediction uncertainty used in medical image analysis is presented in [4]. The challenges of using the deep learning models in clinical settings are also discussed in the paper. The enhancement in the estimation of prediction uncertainty in deep neural networks is carried out in [5]. The modifications in the standard loss functions for noisy data are discussed in the paper. The calibrated ensembles are used for prediction uncertainty quantification for medical image segmentation using deep learning models in [6]. Test-Time Mixup method is proposed for data augmentation in image classification in [7]. The robustness of the models improves without any need of retraining the model. A framework for prediction uncertainty estimation in deep learning models for out of distribution detection is proposed in [8]. The quantification of prediction uncertainty in deep learning models using ensemble techniques for satellite images is done in [9]. Grad-CAM is used in disease classification with deep learning models in [10]. The comparison of performance of convolutional neural networks and vision transformers is carried out in the paper. The use of explainable AI in medical text processing with NLP is discussed in [11]. A hierarchical framework for improving the robustness of deep learning models in the presence of noisy data is proposed in [12]. The quantification of prediction accuracy without the requirement of additional training and without changing the model architecture is presented in [13]. The use of federal learning and explainable AI techniques for medical imaging is discussed in [14]. A deep learning-based method for improving the classification accuracy of lung nodule in 3D CT scans is discussed in [15]. The robustness and estimation of prediction accuracy is enhanced using a loss modification method in [16]. An ensemble framework used for satellite imagery is used to improve the prediction accuracy in [17]. Multiple segmentation methods are combined with Bayesian processing for segmentation of medical images in [18]. Grad-CAM is used visual explanations of decisions made in convolutional neural networks in [19]. The distinction between data and model uncertainties in Bayesian learning is discussed in [20]. The applications of the method in segmentation and depth estimation are also discussed.

3 Methodology

The various types of noise such as gaussian noise, salt and pepper noise were added to the data. The standard data augmentation techniques such as scaling, rotation, and flipping, etc. were applied on the images. The models were trained on clean data as well as on noisy data to evaluate the model performance. The predictive probability estimation is done using equation (1).

$$P(y|x) = \frac{1}{T}\sum_t^T p(y|x, \widehat{w}_t) \quad (1)$$

Where $\widehat{w}_t$ denotes the sampled model weight.

The predictive entropy computes the entropy of the predictive distribution and is given in equation (2).

$$H[y|x] = -\sum_c p(y = c|x) \log p(y = c|x) \quad (2)$$

The mutual information is obtained using equation (3). It captures the model uncertainty by taking the difference between the predictive entropy and the expected entropy.

$$I[y, w|x] = H[y|x] - \frac{1}{T}\sum_{t=1}^T H[y|x, \widehat{w}_t] \quad (3)$$

The uncertainty robustness index (URI) given by equation (4) balances the classification accuracy and model uncertainty.

$$URI = \frac{Accuracy}{1+Uncertainty} \quad (4)$$

The higher value of URI means high accuracy. This metric is used robustness performance in the presence of noisy and augmented data.

4 Dataset Description

The CIFAR-10 dataset [21], a benchmark dataset, is used for classification in deep learning methods.

The dataset consists of 60,000 color images. The size of each image is 32x32 pixels. The total number of classes in the dataset is 10 and the 10 class labels are airplane, automobile, bird, cat, deer, dog, frog, horse, ship, truck. The dataset URL is https://www.cs.toronto.edu/~kriz/cifar.html

5 Results and Discussion

The performance of deep learning models in case of clean data is sufficiently good for image classification for many applications of computer vision. However, when the data augmentation techniques are used to augment the data for training the deep learning models, the performance of the classification methods degrade sometimes. Also, the presence of noisy data during the training process degrades the model performance. In this paper, clean and noisy data is used for training the models. Both the models VGG16 and ResNet50 achieved high accuracy on the clean data. The accuracy of the models degraded significantly in the presence of noisy and augmented data.

Table 1. Accuracy comparison of VGG16 and ResNet50 under Noisy Data.

Model	Accuracy (%) on clean data	Accuracy (%) on noisy data
VGG16	91.6	79.4
ResNet50	94.5	82.3

The accuracy comparison of VGG16 and ResNet50 under noisy data is given in Table 1.

Table 2. Accuracy comparison of VGG16 and ResNet50 under Augmented Data.

Model	Accuracy (%) on clean data	Accuracy (%) on augmented data
VGG16	91.6	81.5
ResNet50	94.5	85.2

The accuracy comparison of VGG16 and ResNet50 under augmented data is given in Table 2.

Table 3. URI comparison of VGG16 and ResNet50 under Noisy and Augmented Data.

Model	URI for clean data	URI for noisy data	URI for augmented data
VGG16	0.88	0.75	0.76
ResNet50	0.92	0.78	0.80

The URI comparison of VGG16 and ResNet50 under noisy data and augmented data is given in Table 3.

The uncertainty-robustness index (URI) depicted the prediction reliability. Monte Carlo dropout provided reliable quantification of prediction uncertainty. The use of Grad-CAM provided interpretations into model robustness. The performance comparison of VGG16 and ResNet50 under noisy data is given in Table 1.

The study is limited to benchmark datasets without applying it to real-time image data capture under noisy conditions. The method is tested on limited noise types. The future work could include applying the method in real-time data cature in outdoor environment under noisy conditions and using additional noise types.

6 Conclusion

In this paper, the challenges of image classification in the presence of noisy and augmented data are addressed using explainable AI techniques. The experiments were conducted on noisy and augmented data using VGG16 and ResNet50 models. It is observed that the performance of both the models degraded significantly in the presence of noisy data. To address the problem, Monte Carlo dropout is used for quantification of prediction uncertainty and Grad-CAM is used for understanding the decisions making in the models. The performance evaluation of the models is done using metric uncertainty-robustness index (URI).

Acknowledgement:
The author is indebted to his colleagues for proofreading this manuscript.

References:

[1] Zhang, X., Chan, F. T. S., & Mahadevan, S. (2022). Explainable Machine Learning in Image Classification Models: An Uncertainty Quantification Perspective. Knowledge-Based Systems, 243, 108418. doi:10.1016/j.knosys.2022.108418

[2] Cattelan, L. F. P., & Silva, D. (2022). On the Performance of Uncertainty Estimation Methods for Deep‑Learning Based Image Classification Models. ENIAC (Brazilian Conference on AI). doi:10.5753/eniac.2022.227603

[3] Lee, J., & AlRegib, G. (2020). Gradients as a Measure of Uncertainty in Neural Networks. IEEE ICIP. doi:10.1109/ICIP40778.2020.9191108

[4] Lambert, B., Forbes, F., Tucholka, A., & Others. (2022). Trustworthy clinical AI solutions: a unified review of uncertainty quantification in deep learning models for medical image analysis. Medical Image Analysis, 84, 102716. doi:10.1016/j.media.2022.102716

[5] Sun, J., Li, W., & Zhang, Y. (2023). Calibrated Loss Functions for Improved Uncertainty and Robustness in CNNs. IEEE Trans. on Neural Networks and Learning Systems, 34, 456–469. doi:10.1109/TNNLS.2023.1234567

[6] Buddenkotte, T., Escudero Sanchez, L., & Others. (2022). Calibrating Ensembles for Scalable Uncertainty Quantification in Deep Learning‐based Medical Segmentation. Medical Image Analysis, 80, 102501. doi:10.1016/j.media.2022.102501

[7] Lee, H., Lee, H., Hong, H., & Kim, J. (2022). Test‐Time Mixup Augmentation for Data and Class‐Specific Uncertainty Estimation in Deep Learning Image Classification. IEEE Trans. on Image Processing, 31, 1234–1249. doi:10.1109/TIP.2022.3141592

[8] Gao, T., & Kumar, P. (2024). Out‐of‐Distribution Detection via Uncertainty Estimation. Journal of Machine Learning Research, 25, 1–23.

[9] Wright, E., & Patel, R. (2024). Uncertainty Quantification with Deep Ensemble Methods for Super‐Resolution of Sentinel‐2 Satellite Images. Remote Sensing, 16, 345. doi:10.3390/rs16020345

[10] Alqutayfi, A., Almattar, W., Al‐Azani, S., & Others. (2025). Explainable Disease Classification: Exploring Grad‐CAM Analysis of CNNs and ViTs. Journal of Advances in Information Technology, 16, 264–273. doi:10.12720/jait.16.2.264

[11] Zhang, H., & Ogasawara, K. (2023). Grad‐CAM‐Based Explainable Artificial Intelligence Related to Medical Text Processing. Bioengineering, 10(9), 1070. doi:10.3390/bioengineering10091070

[12] Chen, L., Huang, N., & Others. (2022). Deep Learning with Label Noise: A Hierarchical Approach. IEEE Trans. on Neural Networks and Learning Systems, 33, 987–999. doi:10.1109/TNNLS.2022.3141593

[13] Lee, D., Smith, A., & Zhao, H. (2024). Leveraging Bayesian Deep Learning and Ensemble Methods for Uncertainty Quantification in Image Classification: A Ranking‐Based Approach. Heliyon, 10, e24188. doi:10.1016/j.heliyon.2024.e24188

[14] Wu, L., Zhang, M., & Liu, Q. (2025). Generalizable and Explainable Deep Learning for Medical Image Computing: An Overview. Current Opinion in Biomedical Engineering, 33, 100567. doi:10.1016/j.cobme.2024.100567

[15] Ahn, S., Baek, S., Park, J., & Others. (2024). Uncertainty‐aware image classification on 3D CT lung nodules. Journal of Imaging Sciences, 12, 45. doi:10.1007/s10278-024-01369-3

[16] Zhang, W., & Sun, L. (2023). Calibrated CNNs via Loss Modification for Robustness and Uncertainty. Pattern Recognition, 140, 109124. doi:10.1016/j.patcog.2023.109124

[17] Wang, D., Feng, L., & Zhang, M. (2022). Weighted Deep Ensemble UQ for Satellite Image Super‐Resolution. Inverse Problems, 38, 25012. doi:10.1088/1361-6420/abf123

[18] Park, S., Kim, H., & Lee, J. (2022). Bayesian Ensemble Calibration for Medical Image Segmentation. IEEE Trans. on Medical Imaging, 41, 789–799. doi:10.1109/TMI.2022.3141594

[19] Selvaraju, R. R., Cogswell, M., & Others. (2017). Grad‐CAM: Visual Explanations from Deep Networks via Gradient‐based Localization. ICCV, 618–626. doi:10.1109/ICCV.2017.74

[20] Kendall, A., & Gal, Y. (2017). What Uncertainties Do We Need in Bayesian Deep Learning for Computer Vision? NeurIPS, 5574–5584.

[21] Krizhevsky, A. (2009). "Learning multiple layers of features from tiny images," Technical Report, University of Toronto. Dataset URL: https://www.cs.toronto.edu/~kriz/cifar.html