Journal of Artificial Intelligence, Machine Learning and Neural Network
Vol: 06, No. 02, July-Dec 2026
ISSN: 2799-1172, DOI: 10.55529/jaimlnn.62.32.43

Research Paper

Same accuracy, different failures: Robustness of pixel-based, gradient-feature and augmented classifiers to common image corruptions

Bilal Öztürk[1*], Kelvin Lleins Rojas-Córdova[2]
[1*]Department of Software Engineering, Faculty of Engineering, Istanbul Aydin University, Istanbul, Turkey.
[2]Escuela Profesional de Ingeniería de Sistemas e Informática, Facultad de Ingeniería de Sistemas e Informática, Universidad Nacional de San Martín, Tarapoto, Peru.

Article Info

Article History:
Received: 21 June 2026
Revised: 20 August 2026
Accepted: 12 September 2026
Published: 22 October 2026

Keywords:
Image Classification
Corruption Robustness
Data Augmentation
Histogram of Oriented Gradients
Distribution Shift
Fashion-MNIST

ABSTRACT

Image classifiers are usually compared by their accuracy on clean test images, although deployed systems receive images degraded by noise, blur, poor contrast, compression and misalignment. This study examines whether clean accuracy predicts robustness to such corruptions and how two common remedies, hand-crafted gradient features and data augmentation, change the picture. Seven classifiers were trained on 20,000 real product photographs from the Fashion-MNIST dataset: logistic regression, random forest and multilayer perceptron (MLP) models on raw pixels; the same three models on histogram-of-oriented-gradients (HOG) features; and an MLP trained with noise and blur augmentation. Each was tested on 5,000 held-out images, both clean and under nine corruptions at three severities, with three training seeds. Clean accuracy varied only between 82.8% and 87.0%, whereas mean corrupted accuracy ranged from 48.6% to 71.0%. The pixel MLP and the random forest differed by 0.7 points on clean images but by 5.0 points under corruption, and the corruption to which each model was most vulnerable differed between models. Contrast reduction reduced the accuracy of pixel models to 12.6–34.3%, whereas HOG models retained 83.4–85.2%, because block normalisation makes HOG descriptors invariant to contrast scaling; the same features, however, were highly sensitive to noise. Augmentation with noise and blur raised the MLP's mean corrupted accuracy from 64.2% to 71.0% without reducing clean accuracy, and it also improved robustness to related unseen corruptions (salt-and-pepper noise and motion blur) and, most strikingly, to contrast reduction (from 12.4% to 64.5% at the highest severity). No model was robust to rotation or translation, and augmentation slightly reduced accuracy under both. Clean accuracy is therefore an unreliable guide to robustness, and robustness should be evaluated separately for each type of corruption expected in deployment.

Corresponding Author:

Journal of Artificial Intelligence, Machine Learning and Neural Network (JAIMLNN)

Bilal Öztürk
Department of Software Engineering, Faculty of Engineering, Istanbul Aydin University, Istanbul, Turkey.
Email: bilalo@aydin.edu.tr

Copyright © 2026 The Author(s). This is an open access article distributed under the Creative Commons Attribution License, (http://creativecommons.org/licenses/by/4.0/) which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

1. INTRODUCTION

Image classifiers are almost always compared by their accuracy on a held-out set of clean images drawn from the same distribution as the training data [1], [2]. Images encountered in deployment, however, are routinely degraded: sensors add noise, camera or object motion causes blur, poor lighting lowers contrast, compression introduces block artefacts, and objects appear rotated, shifted or partly occluded. A classifier's accuracy under such conditions can differ substantially from its clean accuracy [3], [4], [5], [6]. Hendrycks and Dietterich [7] formalised this concern with benchmarks of common corruptions, and Mu and Gilmer [8] constructed an analogous benchmark for handwritten digits.

Two remedies are commonly proposed. The first, older one is to use hand-crafted features designed to be invariant to nuisance factors; the histogram of oriented gradients (HOG) [9], for example, normalises local gradient histograms and was designed to be insensitive to illumination changes, in the tradition of descriptors such as SIFT [10]. The second is data augmentation [11], in which training images are deliberately corrupted so that the model learns to ignore the corruption; methods such as AugMix [12] and noise-based training [13] have improved corruption robustness in deep networks. It is known that robustness gained by augmentation does not always transfer to corruptions not seen during training [6], [7], and that robustness and clean accuracy are often correlated across models [14], [15].

This paper examines these questions with a controlled, reproducible experiment on real product photographs from the Fashion-MNIST dataset [2]. Seven classifiers of three families, trained on raw pixels, on HOG features, or with augmentation, are evaluated under nine corruption types at three severities. The contributions are as follows:

A quantitative comparison showing that classifiers with nearly identical clean accuracy differ by up to 22 percentage points in mean accuracy under corruption, and that each model has a different worst-case corruption.

A demonstration, using real images and their HOG descriptors, that gradient features are exactly invariant to contrast reduction but highly sensitive to noise, and that this trade-off is directly reflected in classification accuracy.

Evidence that noise-and-blur augmentation improves robustness to both seen and several unseen corruptions without reducing clean accuracy, but does not help, and slightly hurts, under geometric transformations.

A per-class analysis identifying which garment categories are most vulnerable to corruption.

2. RELATED WORK

2.1 Corruption Robustness of Image Classifiers

Dodge and Karam [3] showed that deep networks are particularly sensitive to blur and noise, and [4] and [5] studied the effect of blur and other distortions on convolutional networks. [6] Compared humans and deep networks under distortions and found that networks trained on one distortion generalised poorly to others; they later showed that ImageNet-trained networks rely heavily on texture [16]. Hendrycks and Dietterich [7] introduced ImageNet-C with 15 corruption types at five severities, and

Mu and Gilmer [8] introduced MNIST-C. [17] Related robustness to Gaussian noise to adversarial vulnerability [18], [19]. At the level of natural distribution shift, [20], [15] and [14] showed that out-of-distribution accuracy is often strongly correlated with in-distribution accuracy across models; dataset shift in general is reviewed [21].

2.2 Invariant Features and Data Augmentation

Hand-crafted descriptors were designed with invariance in mind. SIFT [10] normalises descriptors for robustness to illumination change, and HOG [9] contrast-normalises gradient histograms over overlapping blocks. Data augmentation instead teaches invariance from data [11]. [12] Proposed AugMix, which mixes chains of augmentations to improve robustness to unseen corruptions, and [13] showed that training with Gaussian and adversarially learned noise improves robustness across many corruptions. Most of these studies concern large convolutional networks on natural-image benchmarks; the behaviour of simpler models and of classical invariant features under the same corruptions has received less direct comparison.

3. METHODOLOGY

3.1 Data

Fashion-MNIST [2] consists of 70,000 grey-scale product photographs of clothing and accessories from an online retailer, reduced to 28 × 28 pixels, in ten classes: T-shirt/top, trouser, pullover, dress, coat, sandal, shirt, sneaker, bag and ankle boot (Figure 1). It was designed as a more challenging drop-in replacement for MNIST [1]. Pixel intensities were scaled to [0, 1]. For each of three random seeds, a stratified sample of 20,000 of the 60,000 official training images (2,000 per class) was used for training. A fixed stratified sample of 5,000 of the 10,000 official test images (500 per class) was used for all evaluations.



Figure 1. Examples of the Real Product Photographs in the Fashion-MNIST Test Set, Four per Class (Displayed with Inverted Grey Scale, so that Darker Pixels Correspond to Higher Intensity)

3.2 Corruptions

Nine corruption types, similar to those of ImageNet-C and MNIST-C [7], [8], were applied to the test images at three severities (Table 1 and Figure 2): additive Gaussian noise, salt-and-pepper noise, Gaussian blur, horizontal motion blur, contrast reduction towards the image mean, rotation, translation, occlusion by a black square at a random position, and JPEG compression [22]. All corrupted images were clipped to [0, 1]. Random elements (noise, occlusion position, rotation and translation direction) were generated with fixed seeds, so that every model was evaluated on exactly the same 135,000 corrupted images (5,000 images × 9 corruptions × 3 severities).

Table 1. Corruption Types and Severity Parameters

Corruption	Parameter	Severity 1	Severity 2	Severity 3
Gaussian noise	Noise standard deviation	0.08	0.16	0.28
Salt & pepper	Fraction of pixels set to 0 or 1	0.03	0.08	0.16
Gaussian blur	Kernel standard deviation (pixels)	0.6	1.0	1.6
Motion blur	Horizontal kernel length (pixels)	3	5	7
Contrast	Contrast factor	0.6	0.4	0.2
Rotation	Angle (degrees, random sign)	10	20	35
Translation	Shift in each axis (pixels, random sign)	2	3	5
Occlusion	Side of black square (pixels)	6	10	14
JPEG	Quality factor	30	15	6

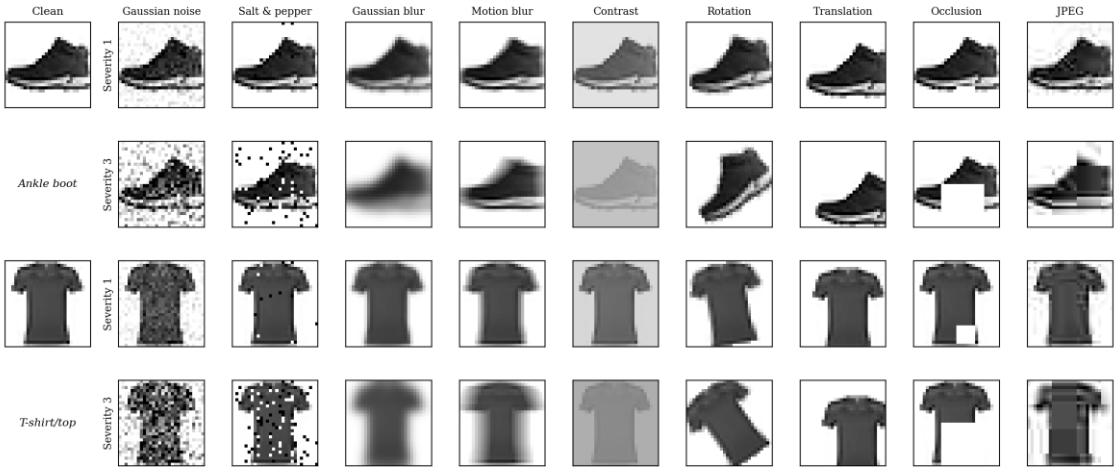


Figure 2. Two Test Images (Ankle Boot and T-Shirt/Top) under the Nine Corruptions at Severity 1 (Upper Row of Each Pair) and Severity 3 (Lower Row). Images are Displayed with Inverted Grey Scale; the Occluding Square therefore Appears White

3.3 Features and Classifiers

Two input representations were used. Pixel models received the 784 raw intensities. HOG models received a 324-dimensional histogram-of-oriented-gradients descriptor [9] with nine orientations, 7 × 7-pixel cells and 2 × 2-cell blocks with L2-Hys block normalisation, computed with scikit-image [23]. Three classifiers were applied to each representation using scikit-learn [24]: multinomial logistic regression (LR) [25] on standardised inputs (C = 0.1); a random forest (RF) [26] with 200 trees; and a multilayer perceptron (MLP) [27] with hidden layers of 256 and 128 units, L2 penalty 10^{-4} and Adam [28] with early stopping. A seventh model, MLP (pixels) + augmentation, used the same architecture but was trained on images of which a random half had been corrupted with either Gaussian noise (standard deviation drawn uniformly from 0.04 to 0.30) or Gaussian blur (kernel standard deviation from 0.4 to 1.8), with equal probability. Noise and blur were therefore "seen" corruptions for this model, while the other seven were "unseen". No model saw corrupted test images or tuned any setting on them.

3.4 Evaluation and Statistical Analysis

Accuracy was computed on the clean test sample and on each of the 27 corrupted versions. The mean corrupted accuracy (mCA) of a model is the unweighted mean over the 27 corruption–severity combinations, and relative robustness is mCA divided by clean accuracy. All results were averaged over the three training seeds, whose standard deviations were small (at most 0.49 points for clean accuracy and 1.65 points for mCA). Because all models were evaluated on the same test images, paired differences

between two models on a given corruption were tested with McNemar's test [29], [30] using the predictions of the first seed. Across-model association between clean accuracy and mCA was measured with Spearman's rank correlation.

4. RESULTS AND DISCUSSION

4.1 Clean Accuracy Does Not Predict Robustness

Table 2 summarises clean and corrupted performance. On clean images, the seven models were separated by only 4.2 percentage points, from 82.8% (LR on pixels) to 87.0% (MLP on pixels). Under corruption, the spread widened to 22.4 points, from 48.6% to 71.0% mean corrupted accuracy. The ranking also changed. The pixel-based random forest was 0.7 points behind the pixel MLP on clean images but 5.0 points behind under corruption, while LR on HOG features was 2.6 points behind the random forest on clean images but 2.8 points ahead under corruption. The augmented MLP was essentially as accurate as the standard MLP on clean images (86.8% versus 87.0%) but the most robust model overall (71.0%), retaining 81.8% of its clean accuracy. Across the seven models, clean accuracy and mCA were positively but imperfectly rank-correlated (Spearman $\rho = 0.75$).

Table 2. Clean Accuracy, Mean Corrupted Accuracy (mCA) and Relative Robustness (mCA / Clean) of the Seven Classifiers (Mean ± SD over Three Training Seeds). Seen Corruptions are Gaussian Noise and Gaussian Blur; the Other Seven are Unseen

Model	Clean Accuracy (%)	mCA (%)	Relative Robustness (%)	Seen Corruptions (%)	Unseen Corruptions (%)
LR (pixels)	82.81 ± 0.17	48.56 ± 1.48	58.6	54.2	47.0
RF (pixels)	86.26 ± 0.36	59.21 ± 0.18	68.6	68.0	56.7
MLP (pixels)	87.00 ± 0.24	64.20 ± 1.65	73.8	78.3	60.2
MLP (pixels) + augmentation	86.76 ± 0.05	70.95 ± 0.04	81.8	84.5	67.1
LR (HOG)	83.64 ± 0.39	62.03 ± 0.11	74.2	58.0	63.2
RF (HOG)	83.35 ± 0.49	60.53 ± 0.13	72.6	48.2	64.1
MLP (HOG)	85.15 ± 0.25	63.72 ± 0.38	74.8	60.7	64.6

4.2 Each Model Fails Differently

Figure 3 and Table 3 show accuracy for each corruption. The profiles differed markedly between model families. Pixel models were devastated by contrast reduction (12.6–34.3%, near chance for LR and RF), whereas all HOG models retained 83.4–85.2%, virtually their clean accuracy. HOG models, in turn, were the most sensitive to Gaussian noise (20.1–44.1%), where the pixel-based MLP and RF retained 68.6–74.3%. The random forest on pixels was the most robust model to salt-and-pepper noise (85.1%), presumably because individual extreme pixels affect only the few splits that use them. JPEG compression and blur were comparatively benign for most models. Rotation and translation were harmful to every model (24.8–51.7% on average and 12.9–30.2% at the highest severity); HOG features, whose cells are tied to fixed image positions, did not confer shift invariance and were only slightly more robust to translation than pixels. The corruption that caused the largest loss was contrast for LR and RF on pixels, translation for the two pixel MLPs, rotation for LR and MLP on HOG, and Gaussian noise for RF on HOG.

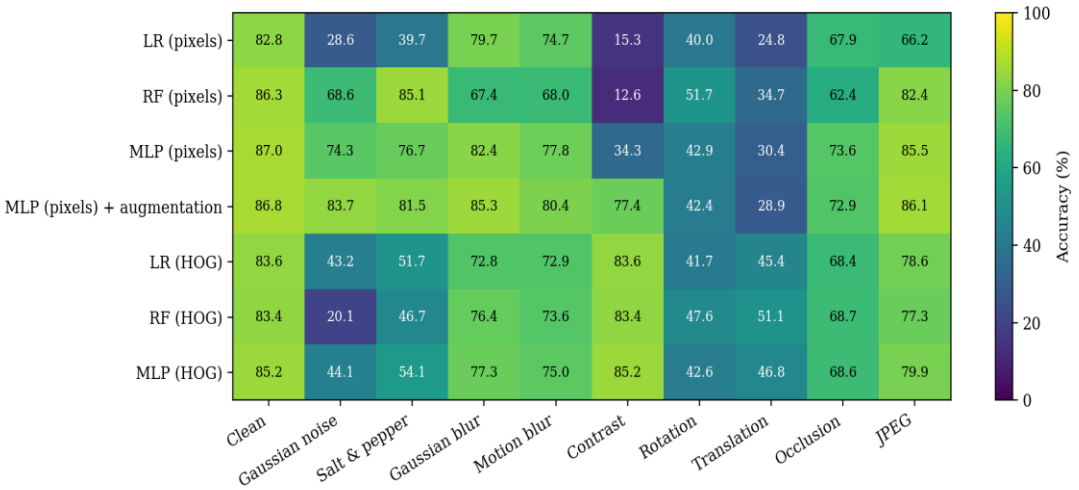


Figure 3. Accuracy (%) of Each Classifier on Clean Test Images and under Each Corruption, Averaged over Three Severities and Three Training Seeds

Table 3. Accuracy (%) under each Corruption, Averaged over Three Severities and Three Seeds. The Most Accurate Model for each Corruption is in Bold

Corruption	LR (Pixels)	RF (Pixels)	MLP (Pixels)	MLP (Pixels) + Aug.	LR (HOG)	RF (HOG)	MLP (HOG)
Gaussian noise	28.6	68.6	74.3	83.7	43.2	20.1	44.1
Salt & pepper	39.7	85.1	76.7	81.5	51.7	46.7	54.1
Gaussian blur	79.7	67.4	82.4	85.3	72.8	76.4	77.3
Motion blur	74.7	68.0	77.8	80.4	72.9	73.6	75.0
Contrast	15.3	12.6	34.3	77.4	83.6	83.4	85.2
Rotation	40.0	51.7	42.9	42.4	41.7	47.6	42.6
Translation	24.8	34.7	30.4	28.9	45.4	51.1	46.8
Occlusion	67.9	62.4	73.6	72.9	68.4	68.7	68.6
JPEG	66.2	82.4	85.5	86.1	78.6	77.3	79.9

4.3 Why HOG Features Are Immune to Contrast but Not to Noise

Figure 4 illustrates the mechanism with a real test image. HOG computes local gradient orientations and normalises each block of histograms to unit norm. Reducing contrast about the image mean multiplies all gradients by the same factor, which the block normalisation removes; the HOG descriptor of the low-contrast sneaker is therefore identical to that of the clean image (correlation 1.00). Additive noise, by contrast, creates spurious gradients in every cell, including the empty background, and the normalisation amplifies them in low-signal regions; the descriptor of the noisy image correlates only 0.40 with the clean one. Translation moves the object relative to the fixed cell grid and reduces the correlation to 0.25. These invariances and sensitivities map directly onto the accuracy pattern in Table 3.

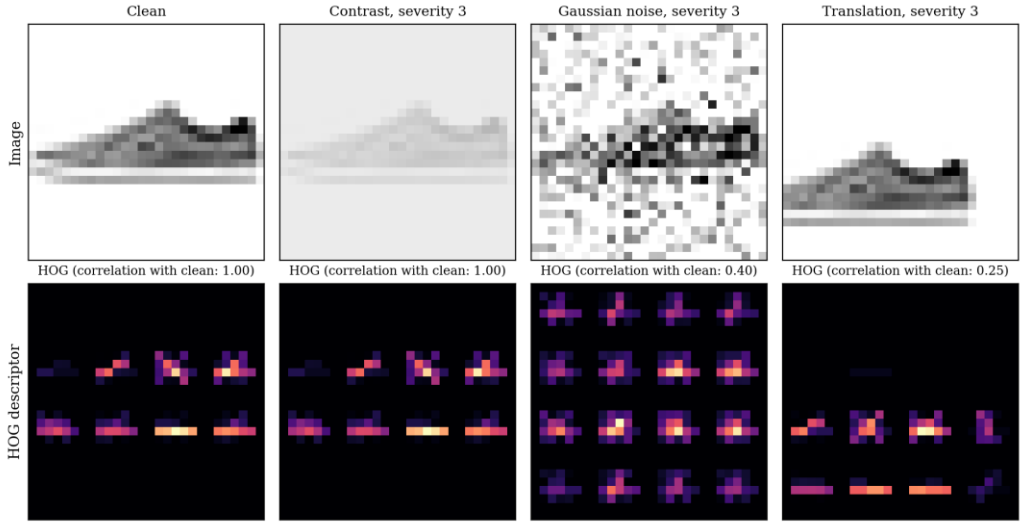


Figure 4. A Real Sneaker Image (Top) and Its HOG Descriptor (Bottom) under Contrast Reduction, Gaussian Noise and Translation at Severity 3. The Correlation with the Clean Descriptor Explains Why HOG-Based Classifiers Are Immune to Contrast Changes but Vulnerable to Noise and Misalignment

4.4 Effect of Augmentation

Training with noise and blur raised the MLP's accuracy on the seen corruptions from 78.3% to 84.5% and on the unseen ones from 60.2% to 67.1%, while clean accuracy changed by only −0.2 points. Table 4 lists the paired differences at the highest severity. The largest gains were on Gaussian noise (+21.0 points) and, unexpectedly, on contrast reduction (+58.4 points), an unseen corruption; gains were also significant for salt-and-pepper noise (+8.4), Gaussian blur (+7.2) and motion blur (+2.7). Training on noisy and blurred images appears to have made the network rely less on absolute intensity levels, although this study did not test the mechanism directly. In contrast, augmentation slightly but significantly reduced accuracy under rotation (−3.0 points), translation (−2.5) and occlusion (−1.5) and had no significant effect under JPEG compression. The augmented model was therefore not uniformly more robust: it traded small losses on geometric corruptions for large gains on photometric ones.

Table 4. Effect of Augmentation at Severity 3: Accuracy of the Augmented MLP Minus that of the Standard Pixel MLP, with McNemar's Test on the 5,000 Paired Test Predictions [31]

Corruption	Seen in Training?	Standard MLP (%)	Augmented MLP (%)	Difference (Points)	Adjusted P
Gaussian noise	Yes	59.3	80.4	+21.0	< 0.001
Gaussian blur	Yes	76.8	83.4	+7.2	< 0.001
Salt & pepper	No	66.6	77.1	+8.4	< 0.001
Motion blur	No	69.2	73.8	+2.7	< 0.001
Contrast	No	12.4	64.5	+58.4	< 0.001
Rotation	No	22.3	19.3	−3.0	< 0.001
Translation	No	15.1	13.6	−2.5	< 0.001
Occlusion	No	61.6	60.1	−1.5	0.003
JPEG	No	84.4	85.3	+0.7	0.082

Accuracies in the table are means over three seeds; the differences and tests use the paired predictions of the first seed, so a difference can deviate slightly from the difference of the displayed means.

Figure 5 shows how accuracy decreased with severity. All models lost accuracy steadily, but the augmented MLP degraded most slowly, retaining 61.9% at the highest severity compared with 52.0% for

the standard MLP and 39.2% for LR on pixels. At severity 1, the HOG-based MLP was nearly as robust as the augmented MLP (77.7% versus 79.5%), but its advantage over pixel models largely disappeared at severity 3 because of its sensitivity to strong noise.

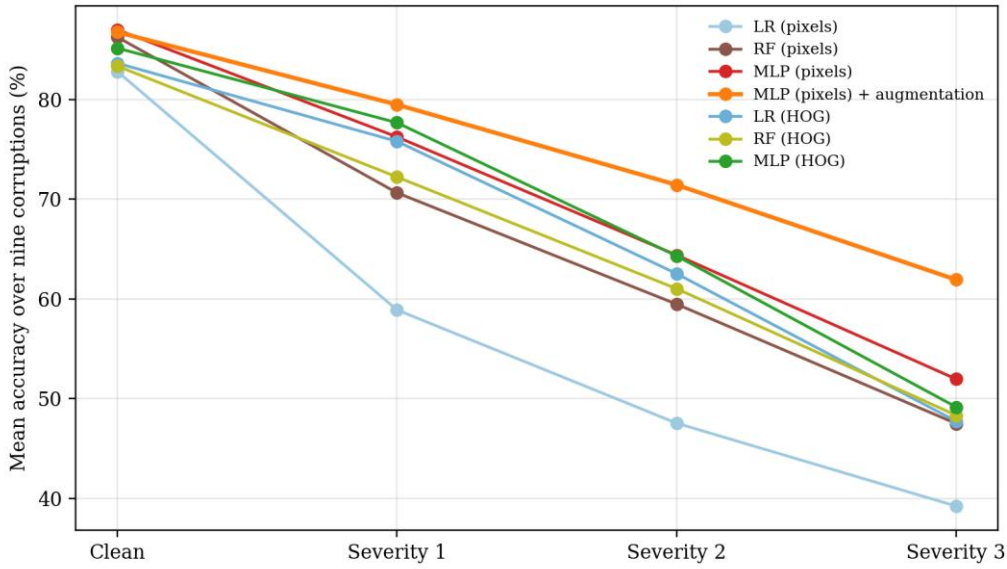


Figure 5. Mean Accuracy over the Nine Corruptions as a Function of Severity (Severity 0 = Clean Images), Averaged over Three Training Seeds

4.5 Which Classes Are Most Vulnerable?

Figure 6 shows the recall of each class averaged over all corruptions. For the standard pixel MLP, shirts were the most vulnerable category (37.0%), followed by coats (50.5%) and pullovers (54.3%), all of which are visually similar upper-body garments whose distinguishing details (collars, buttons, textures) are easily destroyed by noise or blur. Bags were almost unaffected (94.6%), because their distinctive overall shape survives most corruptions. Augmentation improved the recall of the most vulnerable classes most, raising shirts to 54.8% and coats to 62.7%, whereas the HOG-based MLP helped T-shirts and sandals but not pullovers or coats.

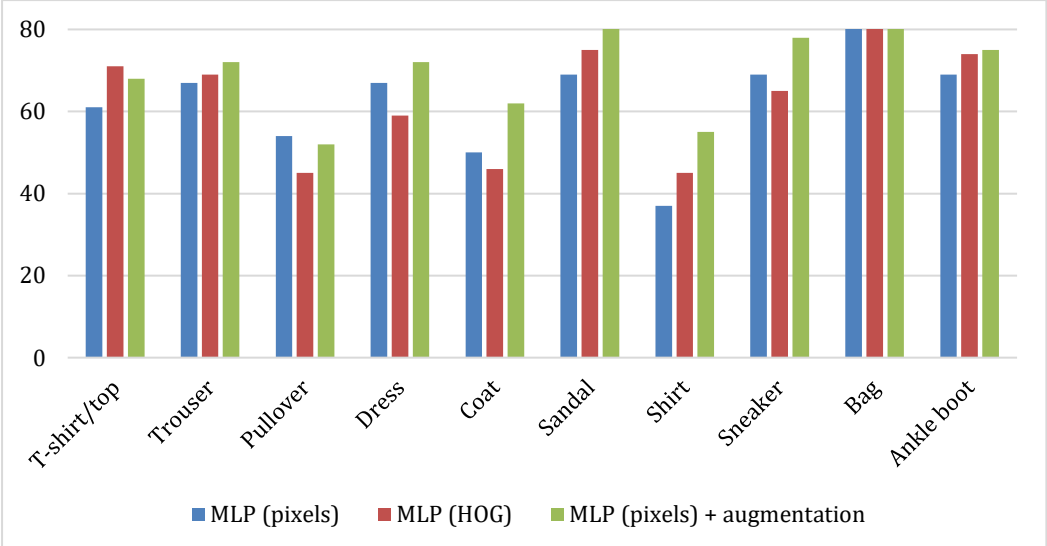


Figure 6. Recall of Each Class under Corruption (Averaged over Nine Corruptions and Three Severities, First Seed) for the Standard Pixel MLP, the HOG-Based MLP and the Augmented Pixel MLP

4.6 Discussion

4.6.1 Principal Findings

Clean test accuracy, the figure reported in almost every image-classification study, was a poor guide to how the seven models behaved under realistic image degradations. Models separated by less than one point on clean images differed by five points under corruption, and the model with the best clean accuracy was not the most robust. The failure modes were specific to the representation: pixel models failed under contrast reduction, gradient features failed under noise, and all models failed under geometric transformation. These observations are consistent with the correlation between in- and out-of-distribution accuracy reported for natural shifts [14], [15], which was also present here ($\rho = 0.75$), but they show that this correlation is too weak to rank models reliably for a specific deployment condition.

4.6.2 Invariance Must Match the Threat

The contrast result is a clear example of an invariance that was designed into a representation and that works exactly as intended [9]: HOG classifiers were unaffected by a corruption that reduced pixel classifiers to chance. The same normalisation that provides this invariance, however, amplifies noise in low-signal regions, which made HOG the worst representation under noise. Invariance is therefore not a general-purpose property but a commitment to ignore a particular transformation, and it can increase sensitivity to others. The same holds for augmentation: training on noise and blur protected against photometric degradations, including some unseen ones, but offered no protection against rotation or translation, consistent with earlier findings that robustness transfers only partially across corruption types [6], [7].

4.6.3 Implications for Practice

Four recommendations follow. First, image classifiers intended for deployment should be evaluated under the specific corruptions expected in their operating environment, and results should be reported per corruption rather than only as an average [7], [8]. Second, simple photometric augmentation with noise and blur is inexpensive, did not reduce clean accuracy, and substantially improved robustness to several corruptions; it is a sensible default. Third, geometric robustness requires geometric augmentation or architectures with built-in translation equivariance, such as convolutional networks [1]; neither pixel MLPs nor HOG features provided it. Fourth, when illumination or contrast varies strongly, contrast-normalised inputs or features are an effective, training-free safeguard.

4.6.4 Limitations

The study used small 28 × 28 grey-scale images and relatively simple classifiers; convolutional networks, larger images and colour data may show different robustness profiles, although several of the mechanisms identified here, such as the contrast invariance of normalised gradient features, are general. The corruptions were synthetic, applied at three severities chosen by the authors, and do not cover all real-world degradations. Only one augmentation scheme was examined. Three training seeds were used, and the McNemar tests were computed on a single seed; the small seed-to-seed variability suggests that the conclusions are stable, but formal inference across seeds was not performed. Finally, the observed transfer of augmentation benefits to contrast reduction was not explained mechanistically and warrants further investigation.

5. CONCLUSION

This study evaluated seven classifiers on real Fashion-MNIST product photographs under nine common corruptions at three severities. Clean accuracy varied by only 4.2 percentage points, but mean corrupted accuracy varied by 22.4 points, and each model family had a characteristic failure mode: pixel models collapsed under contrast reduction, HOG models under noise, and all models under rotation and translation. Contrast-normalised HOG features were exactly invariant to contrast changes, as shown by their descriptors on real images, but this invariance came at the cost of noise sensitivity. Augmentation

with noise and blur raised mean corrupted accuracy from 64.2% to 71.0% without reducing clean accuracy and transferred to some unseen corruptions, but not to geometric ones. Clean accuracy is therefore insufficient for selecting an image classifier for deployment; robustness should be measured for each expected type of corruption, and invariances should be chosen to match the degradations that the system will actually face. Future work will extend the analysis to convolutional networks, colour images, real rather than synthetic degradations, and a wider range of augmentation strategies.

Acknowledgments

The authors thank Zalando Research for making the Fashion-MNIST dataset publicly available, the developers of scikit-learn and scikit-image, and the researchers whose work is cited in this article.

Funding Information

This research has received no external funding.

Author Contribution Statement

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Bilal Öztürk	✓	✓	✓		✓				✓		✓			
Kelvin Lleins Rojas-Córdova		✓		✓		✓		✓		✓		✓		

C : **C**onceptualization
M : **M**ethodology
So : **So**ftware
Va : **Va**lidation
Fo : **Fo**rmal analysis
I : **I**nvestigation
R : **R**esources
D : **D**ata Curation
O : Writing - **O**riginal Draft
E : Writing - Review & **E**diting
Vi : **Vi**sualization
Su : **Su**pervision
P : **P**roject administration
Fu : **Fu**nding acquisition

Conflict of Interest Statement

No conflict of interest.

Informed Consent

Not Applicable.

Ethical Approval

Not Applicable.

Data Availability

This study draws on the publicly available Fashion-MNIST dataset (MIT licence). No new datasets were created. The analysis code is available from the corresponding author on reasonable request.

REFERENCES

[1] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," Proceedings of the IEEE, vol. 86, no. 11, pp. 2278–2324, 1998. doi.org/10.1109/5.726791

[2] H. Xiao, K. Rasul, and R. Vollgraf, "Fashion-MNIST: A novel image dataset for benchmarking machine learning algorithms," arXiv preprint arXiv:1708.07747, 2017. doi.org/10.48550/arXiv.1708.07747

[3] S. Dodge and L. Karam, "Understanding how image quality affects deep neural networks," in Proc. 8th Int. Conf. Quality of Multimedia Experience, 2016, pp. 1–6. doi.org/10.1109/QoMEX.2016.7498955

[4] I. Vasiljevic, A. Chakrabarti, and G. Shakhnarovich, "Examining the impact of blur on recognition by convolutional networks," arXiv preprint arXiv:1611.05760, 2016. doi.org/10.48550/arXiv.1611.05760

[5] Y. Zhou, S. Song, and N.-M. Cheung, "On classification of distorted images with deep convolutional neural networks," in Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing, 2017, pp. 1213–1217. doi.org/10.1109/ICASSP.2017.7952349

[6] R. Geirhos, C. R. M. Temme, J. Rauber, H. H. Schütt, M. Bethge, and F. A. Wichmann, "Generalisation in humans and deep neural networks," in Advances in Neural Information Processing Systems 31, 2018, pp. 7538–7550. doi.org/10.48550/arXiv.1808.08750

[7] D. Hendrycks and T. Dietterich, "Benchmarking neural network robustness to common corruptions and perturbations," in Proc. 7th Int. Conf. Learning Representations, 2019. doi.org/10.48550/arXiv.1903.12261

[8] N. Mu and J. Gilmer, "MNIST-C: A robustness benchmark for computer vision," arXiv preprint arXiv:1906.02337, 2019. doi.org/10.48550/arXiv.1906.02337

[9] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in Proc. IEEE Conf. Computer Vision and Pattern Recognition, 2005, pp. 886–893. doi.org/10.1109/CVPR.2005.177

[10] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," International Journal of Computer Vision, vol. 60, no. 2, pp. 91–110, 2004. doi.org/10.1023/B:VISI.0000029664.99615.94

[11] C. Shorten and T. M. Khoshgoftaar, "A survey on image data augmentation for deep learning," Journal of Big Data, vol. 6, p. 60, 2019. doi.org/10.1186/s40537-019-0197-0

[12] D. Hendrycks, N. Mu, E. D. Cubuk, B. Zoph, J. Gilmer, and B. Lakshminarayanan, "AugMix: A simple data processing method to improve robustness and uncertainty," in Proc. 8th Int. Conf. Learning Representations, 2020. doi.org/10.48550/arXiv.1912.02781

[13] E. Rusak et al., "A simple way to make neural networks robust against diverse image corruptions," in Computer Vision – ECCV 2020, Lecture Notes in Computer Science, vol. 12348. Cham: Springer, 2020, pp. 53–69. doi.org/10.1007/978-3-030-58580-8_4

[14] J. P. Miller et al., "Accuracy on the line: On the strong correlation between out-of-distribution and in-distribution generalization," in Proc. 38th Int. Conf. Machine Learning, PMLR, vol. 139, 2021, pp. 7721–7735. proceedings.mlr.press/v139/miller21b.html

[15] R. Taori, A. Dave, V. Shankar, N. Carlini, B. Recht, and L. Schmidt, "Measuring robustness to natural distribution shifts in image classification," in Advances in Neural Information Processing Systems 33, 2020, pp. 18583–18599. doi.org/10.48550/arXiv.2007.00644

[16] R. Geirhos, P. Rubisch, C. Michaelis, M. Bethge, F. A. Wichmann, and W. Brendel, "ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness," in Proc. 7th Int. Conf. Learning Representations, 2019. doi.org/10.48550/arXiv.1811.12231

[17] N. Ford, J. Gilmer, N. Carlini, and E. D. Cubuk, "Adversarial examples are a natural consequence of test error in noise," in Proc. 36th Int. Conf. Machine Learning, PMLR, vol. 97, 2019, pp. 2280–2289. proceedings.mlr.press/v97/gilmer19a.html

[18] C. Szegedy et al., "Intriguing properties of neural networks," in Proc. 2nd Int. Conf. Learning Representations, 2014. doi.org/10.48550/arXiv.1312.6199

[19] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," in Proc. 3rd Int. Conf. Learning Representations, 2015. doi.org/10.48550/arXiv.1412.6572

[20] B. Recht, R. Roelofs, L. Schmidt, and V. Shankar, "Do ImageNet classifiers generalize to ImageNet?," in Proc. 36th Int. Conf. Machine Learning, PMLR, vol. 97, 2019, pp. 5389–5400. proceedings.mlr.press/v97/recht19a.html

[21] J. Quiñonero-Candela, M. Sugiyama, A. Schwaighofer, and N. D. Lawrence, Eds., Dataset Shift in Machine Learning. Cambridge, MA: MIT Press, 2009. doi.org/10.7551/mitpress/9780262170055.001.0001

[22] G. K. Wallace, "The JPEG still picture compression standard," IEEE Transactions on Consumer Electronics, vol. 38, no. 1, pp. xviii–xxxiv, 1992. doi.org/10.1109/30.125072

[23] S. van der Walt et al., "scikit-image: Image processing in Python," PeerJ, vol. 2, p. e453, 2014. doi.org/10.7717/peerj.453
[24] F. Pedregosa et al., "Scikit-learn: Machine learning in Python," Journal of Machine Learning Research, vol. 12, pp. 2825–2830, 2011. jmlr.org/papers/v12/pedregosa11a.html
[25] D. R. Cox, "The regression analysis of binary sequences," Journal of the Royal Statistical Society: Series B (Methodological), vol. 20, no. 2, pp. 215–242, 1958. doi.org/10.1111/j.2517-6161.1958.tb00292.x
[26] L. Breiman, "Random forests," Machine Learning, vol. 45, no. 1, pp. 5–32, 2001. doi.org/10.1023/A:1010933404324
[27] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning representations by back-propagating errors," Nature, vol. 323, no. 6088, pp. 533–536, 1986. doi.org/10.1038/323533a0
[28] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in Proc. 3rd Int. Conf. Learning Representations, 2015. doi.org/10.48550/arXiv.1412.6980
[29] Q. McNemar, "Note on the sampling error of the difference between correlated proportions or percentages," Psychometrika, vol. 12, no. 2, pp. 153–157, 1947. doi.org/10.1007/BF02295996
[30] T. G. Dietterich, "Approximate statistical tests for comparing supervised classification learning algorithms," Neural Computation, vol. 10, no. 7, pp. 1895–1923, 1998. doi.org/10.1162/089976698300017197
[31] S. Holm, "A simple sequentially rejective multiple test procedure," Scandinavian Journal of Statistics, vol. 6, no. 2, pp. 65–70, 1979. www.jstor.org/stable/4615733

How to Cite: Bilal Öztürk, Kelvin Lleins Rojas-Córdova (2026). Same accuracy, different failures: Robustness of pixel-based, gradient-feature and augmented classifiers to common image corruptions. Journal of Artificial Intelligence, Machine Learning and Neural Network, 6(2), 32-43. https://doi.org/10.55529/jaimlnn.62.32.43

BIOGRAPHIES OF AUTHORS

	**Bilal Öztürk,** is a member of the Department of Software Engineering, Faculty of Engineering, Istanbul Aydin University, Istanbul, Turkey. His research covers image processing, image transmission and compression, and the software engineering of image-based systems. He is interested in the behaviour of image restoration and classification methods when they meet data unlike their training material, and in the practical evaluation of such methods. Email: bilalo@aydin.edu.tr
	**Kelvin Lleins Rojas-Córdova,** is a researcher in the Escuela Profesional de Ingeniería de Sistemas e Informática, Facultad de Ingeniería de Sistemas e Informática, Universidad Nacional de San Martín, Tarapoto, Peru. His work is in computer vision and image segmentation, with an interest in how image classifiers behave outside the conditions they were trained on. He is concerned with the robustness of visual models to corruption and distribution shift and with the design of controlled benchmarking protocols.