IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Recent Advances In Synthetic Data Have Enabled The Generation Of Artificial Intelligence (AI) generated images

[1](Asst. professor. Atul Taware) (MCA TIT College Bhopal, Rgpv),India
[2](Asst. professor vijay vishwakarma) (EC TIT College Bhopal , Rgpv),India

ABSTRACT

Recent advances in synthetic data have enabled the generation of images with such high quality that human beings cannot distinguish the difference between real-life photographs and Artificial Intelligence (AI) generated images. Given the critical necessity of data reliability and authentication, this article proposes to enhance our ability to recognise AI-generated images through computer vision. Initially, a synthetic dataset is generated that mirrors the ten classes of the already available CIFAR-10 dataset with latent diffusion, providing a contrasting set of images for comparison to real photographs. The model is capable of generating complex visual attributes, such as photorealistic reflections in water. The two sets of data present as a binary classification problem with regard to whether the photograph is real or generated by AI. This study then proposes the use of a Convolutional Neural Network (CNN) to classify the images into two categories; Real or Fake. Following hyperparameter tuning and the training of 36 individual network topologies, the optimal approach could correctly classify the images with 92.98% accuracy. Finally, this study implements explainable AI via Gradient Class Activation Mapping to explore which features within the images are useful for classification. Interpretation reveals interesting concepts within the image, in particular, noting that the actual entity itself does not hold useful information for classification; instead, the model focuses on small visual imperfections in the background of the images. The complete dataset engineered for this study, referred to as the CIFAKE dataset, is made publicly available to the research community for future work.

INDEX TERMS AI-generated images, generative AI, image classification, latent diffusion.

INTRODUCTION

The field of synthetic image generation by Artificial Intelligence (AI) has developed rapidly in recent years, and the ability to detect AI-generated photos has also become a critical necessity to ensure the authenticity of image data. Within recent memory, generative technology often produced images with major visual defects that were noticeable to the human eye, but now we are faced with the possibility of AI models generating high-fidelity and photorealistic images in a matter of seconds. The AI-generated images are now at the quality level needed to compete with humans and win art competitions [1]. Latent Diffusion Models (LDMs), a type of generative
model, have emerged as a powerful tool to generate synthetic imagery [2]. These recent developments have caused a paradigm shift in our understanding of creativity, authenticity and truth. This has led to a situation where consumer-level technology is available that could quite easily be used for the violation of privacy and

to commit fraud. These philosophical and societal implications are at the forefront of the current state of the art, raising fundamental questions about the nature of trustworthiness and reality. Recent technological.

advances have enabled the generation of images with such high quality that human beings cannot tell the difference between a real-life photograph and an image that is no more than a hallucination of an artificial neural network's weights and biases. Generative imagery that is indistinguishable from photographic data raises questions both ontological, those which concern the nature of being, and epistemological, surrounding the theories of methods, validity, and scope. Ontologically, given that humans cannot tell the difference between images from cameras and those generated by AI models such as an Artificial Neural Network, in terms of digital information, *what is real and what is not?* The epistemological reality is that there are serious questions surrounding the reliability of human knowledge and the ethical implications that surround the misuse of these types of technology. The implications suggest that we are in growing need of a system that can aid us in the recognition of real images versus those generated by AI. This study explores the potential of using computer vision to enhance our newfound inability to recognise the difference between real photographs and those that are AI-generated.

Given that there are many years worth of photographic datasets available for image classification, these provide examples for a model of real images. Following the generation of a synthetic equivalent to such data, we will then explore the output of the model before finally implementing methods of differentiation between the two types of image. There are several scientific contributions with multidisciplinary and social implications that arise from this study.

IMAGE CLASSIFICATION

Image classification is an algorithm that predicts a class label given an input image. The learnt features are extracted from the image and processed in order to provide an output, in this case, whether or not the image is real or synthetic. This subsection describes the selected approach to classification. In this study, the Convolutional Neural Network (CNN) [26], [27], [28] is employed to learn from the input

FIGURE 1. Examples of images from the CIFAR-10 image classification dataset

It is the concatenation of two main networks with intermediate operations. These are the convolutional layers and the fully connected layers. The initial convolutional network within the overall model is the CNN, which can be operationally generalised for an image of dimensions *x* and a Although the goal of the network is to use backpropagation to reduce binary cross-entropy loss, this study also notes an extended number of classification metrics. These are the Precision, which is a measure of how many of the predictive positive cases are positive, a metric which allows for the analysis of false-positives:

Precision =True positives /True positives + False positives

The Recall which is a measure of how many positive cases are correctly predicted, which enables analysis of false-negative predictions:

Recall = True positives/ True positives + False negatives

This measure is particularly important in this case, as it is in fraud detection, since a false negative would falsely accuse
the author of generating their image with AI. Finally, the F-1 score is considered:

F1 score = 2 × Precision × Recall/Precision + Recall

which is a unified metric of precision and recall. The dataset that forms the classification is the collection of real images and the equivalent synthetic images generated, detailed in Sections III-A and III-B, respectively. 100, 000 images are used for training (50, 000 real images and 50, 000 synthetic images), and 20, 000 are used for testing (10, 000 real and 10, 000 synthetic). Initially, CNN architectures are benchmarked as a lone feature extractor. That is, the filters of {16, 32, 64, 128} are benchmarked in layers of {1, 2, 3}, flattened, and Although the goal of the network is to use backpropagation to reduce binary cross-entropy loss, this study also notes an extended number of classification metrics. These are the Precision, which is a measure of how many of the predictive positive cases are positive, a metric which allows for the analysis of false-positives:

Precision = True positives/ True positives + False positives

.

The Recall which is a measure of how many positive cases are correctly predicted, which enables analysis of false-negative predictions:

Recall =True positives/True positives + False negatives

This measure is particularly important in this case, as it is in fraud detection, since a false negative would falsely accuse the author of generating their image with AI. Finally, the F-1 score is considered:

F1 score = 2 × Precision × Recall/Precision + Recall

which is a unified metric of precision and recall. The dataset that forms the classification is the collection of real images and the equivalent synthetic images generated, detailed in Sections III-A and III-B, respectively. 100, 000 images are used for training (50, 000 real images and 50, 000 synthetic images), and 20, 000 are used for testing (10, 000 real and 10, 000 synthetic). Initially, CNN architectures are benchmarked as a lone feature extractor. That is, the filters of {16, 32, 64, 128} are benchmarked in layers of {1, 2, 3}, flattened, and

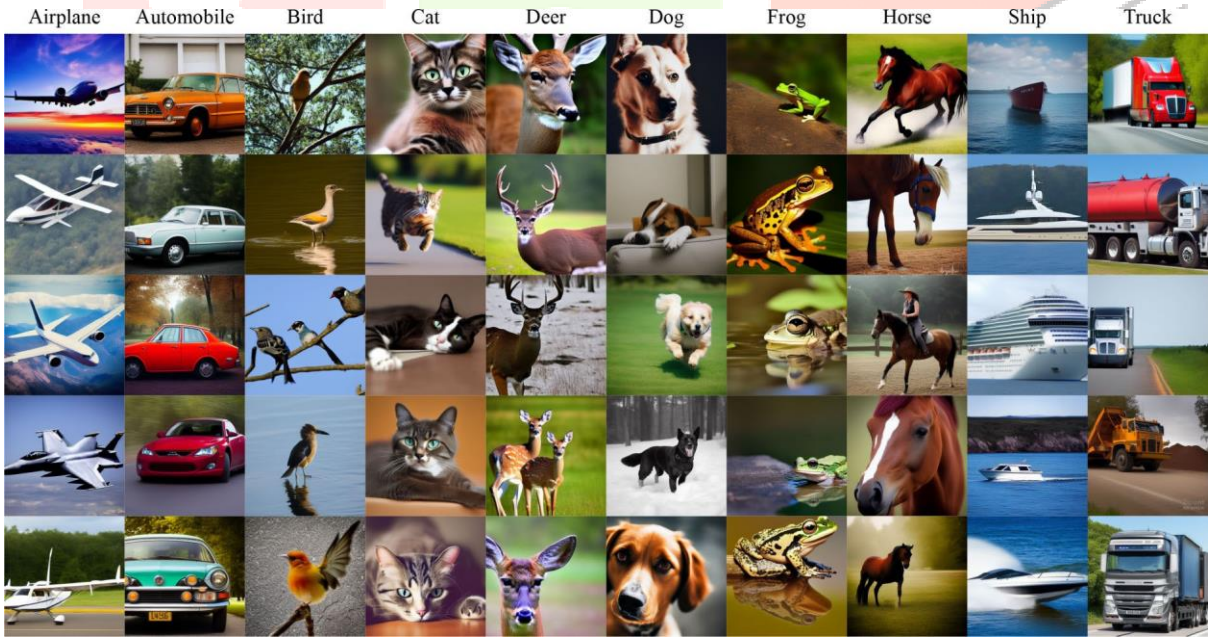

FIGURE 2. Examples of AI-generated images within the dataset contributed by this study, selected at random with regards to their real CIFAR-10 equivalent labels

EXPERIMENTAL HARDWARE AND SOFTWARE

The neural networks used for the detection of AI-generated images were engineered with the TensorFlow library [31]. All TensorFlow seeds were set to 1 for replicability. The Latent Diffusion model used for the generation of synthetic data was Stable Diffusion version 1.4 [2]; Random seed vectors were denoised for a total of 50 steps to form images and the Euler Ancestral scheduler was used. Synthetic images were rendered at a resolution of 512px before resizing to 32px by bilinear interpolation to match the resolution of CIFAR-10. All algorithms in this study were executed using a single Nvidia RTX 3080Ti GPU, which has 10,240 CUDA cores, a clock speed of 1.67 GHz, and 12GB GDDR6X VRAM.

RESULTS AND OBSERVATIONS

This section presents examples of the dataset followed by the findings of the planned computer vision experiments. The dataset is also released to the public research community for use in future studies, given the important implications of detecting AI-generated imagery

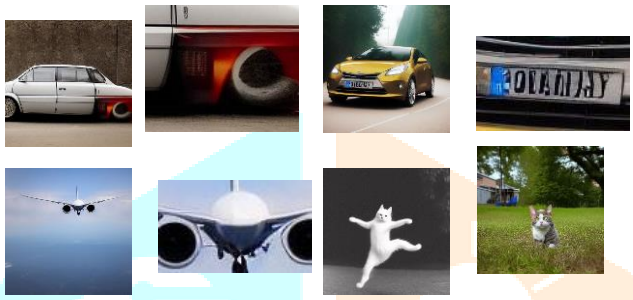

FIGURE 3. Examples of visual defects found within the synthetic image dataset.

TABLE 1. Observed classification accuracy metrics for feature extraction networks.

S.N.	Filters	Layers 1	Layers 2	Layers 3
1	16	90.06	91.46	91.63
2	32	90.38	92.93	92.54
3	64	90.94	92.71	92.38
4	128	90.39	92.98	92.07

DATASET EXPLORATION

Random samples of images used in this study and within the dataset provided can be observed in Figure 2. Five images are presented for each class label, and all of the images within this figure are synthetic, which have been generated by the SDM. Note within this sample that the images are highquality and, for the most part, seem to be difficult to discern as synthetic by the human eye. Synthetic photographs are representative of their counterparts from reality and feature complex attributes such as depth of field, reflections, and motion blur. It can also be observed that there are visual imperfections within some of the images. Figure 3 shows a number of examples of the win of the dataset in which the model has output images with visual glitches. Given that the LAION dataset provides physical descriptions of the image content, little to no information on text is provided, and thus it can be seen that the model produces shapes similar to alphabetic characters. Also observed here is a lack of important detail, such as the case of a jet aircraft that has no cockpit window. It seems that this image has been produced by combining the knowledge of jet aircraft (in particular, the engines) along with the concept of an Unmanned Aerial Vehicle's chassis. Finally, there are also some cases of anatomical errors for living creatures, seen in these examples through the cat's limbs and eyes. Complex visual concepts are present within much of the dataset, with examples shown in Figure 4. Observe that the ripples in the water and reflections of the entities are highly realistic and match what would be expected within a photograph. In addition to complex lighting, there is also evidence of depth of field and photographic framing.

CLASSIFICATION RESULTS

In this subsection, we present the results for the computer vision experiments for image classification. The problem

TABLE 2. Observed validation loss for the filters within the convolutional neural network.

S.N.	Filters	Layers 1	Layers 2	Layers 3
1	16	0.254	0.222	0.21
2	32	0.237	0.18	0.193
3	64	0.226	0.196	0.219
4	128	0.234	0.221	0.259

TABLE 3. Observed validation precision for the filters within the convolutional neural network.

S.N.	Filters	Layers 1	Layers 2	Layers 3
1	16	0.903	0.941	0.921
2	32	0.878	0.923	0.937
3	64	0.908	0.947	0.936
4	128	0.92	0.948	0.94

TABLE 4. Observed validation recall for the filters within the convolutional neural network.

S.N.	Filters	Layers 1	Layers 2	Layers 3
1	16	0.897	0.885	0.911
2	32	0.938	0.936	0.912
3	64	0.92	0.904	0.91
4	128	0.906	0.909	0.898

TABLE 5. Observed validation F1-Score for the filters within the convolutional neural network

S.N.	Filters	Layers 1	Layers 2	Layers 3
1	16	0.9	0.912	0.916
2	32	0.907	0.93	0.924
3	64	0.91	0.925	0.923
4	128	0.913	0.928	0.919

faced by the CNN is that of binary classification, whether or not the image is a real photograph or the output of an LDM. The validation accuracy of the results and the loss metrics for the feature extractors can be found in Tables 2 and 3, respectively. All feature extractors scored relatively well without the need for dense layers to process feature maps, with an average classification accuracy of 91.79%. The lowest loss feature extractor was found to use two layers of 32 filters, which led to an overall classification accuracy of 92.93% and a binary cross-entropy loss of 0.18. The highest accuracy model, two layers of 128 filters, scored 92.98% with a loss of 0.221. Extended validation metrics are presented in Tables 4, 5, and 6, which detail validation precision, recall, and F1 scores, respectively. The F1 score, which is a unification of precision and recall, had a mean value of 0.929 with the highest being 0.936. A small standard deviation of 0.003 was observed. Following these experiments, the lowest-loss feature extractor is selected for further engineering of the network topology. This was the model that had two layers of 32 convolutional filters.

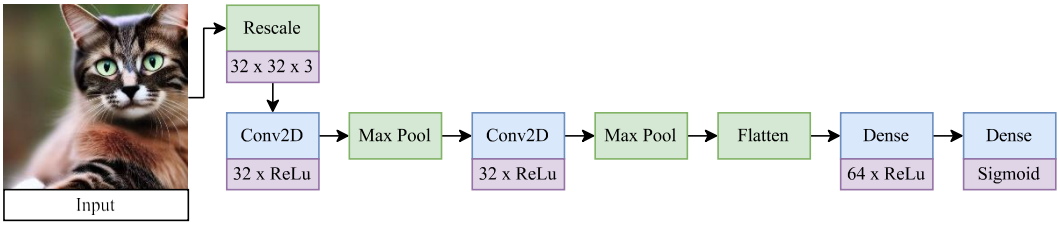

FIGURE 4. An example of one of the final model architectures following hyperparameter search for the classification ofreal or AI-generated images.

The XAI approach also shows an interesting mechanic in a more general sense. Given the examples of airplane, bird, frog, horse, and ship, note that the object within the image has little to no class activation overlay whatsoever. This suggests that the actual focus of the image itself, the entity, contains almost no useful features for synthetic image recognition. This suggests that the model is often available to produce a near-perfect representation of the entity.

CONCLUSION AND FUTURE WORK

This study has proposed a method to improve our waning ability to recognise AI-generated images through the use of Computer Vision and to provide insight into predictions with visual cues. To achieve this, this study proposed the generation of a synthetic dataset with Latent Diffusion, recognition with Convolutional Neural Networks, and interpretation through Gradient Class Activation Mapping. The results showed that the synthetic images were high quality and featured complex visual attributes, and that binary classification could be achieved with around 92.98% accuracy. Grad-CAM interpretation revealed interesting concepts within the images that were useful for predictions. In addition to the method proposed in this study, a significant contribution is made through the release of the CIFAKE dataset. The dataset contains a total of 120, 000 images (60, 000 real images from CIFAR-10 and 60,000 synthetic images generated for this study). The CIFAKE dataset provides the research community with a valuable resource for future work on the social problems faced by AI-generated imagery. The dataset provides a significant expansion of the resource availability for the development and testing of applied computer vision approaches to this problem. The reality of AI generating images that are indistinguishable from real-life photographic images raises fundamental questions about the limits of human perception, and thus this study proposed to enhance that ability by *fighting fire with fire*. The proposed approach addresses the challenges of ensuring the authenticity and trustworthiness of visual data. Future work could involve exploring other techniques to classify the provided dataset. For example, the implementation of attention-based approaches is a promising new field that could provide increased ability and an alternative
method of explainable AI. Furthermore, with even further improvements to synthetic imagery in the future, it is important to consider updating the dataset with images generated by these approaches. Furthermore, considering generating images from other domains, such as human faces and clinical scans, would provide additional datasets for this type of study and expand the applicability of our proposed approach to other fields of research. Finally, in conclusion, this study provides contributions to the ongoing implications of AI-generated images. The proposed approach supports important implications of ensuring data authenticity and trustworthiness, providing not only a system that can recognise synthetic images, but also data and interpretation. The public release of the CIFAKE dataset generated within this study provides a valuable resource for interdisciplinary research.

REFERENCES

[1] K. Roose, ‘‘An AI-generated picturewon an art prize. Artists aren’t happy,’’ *New York Times*, vol. 2, p. 2022, Sep. 2022.
[2] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, ‘‘High-resolution image synthesis with latent diffusion models,’’ in *Proc.*
IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2022, pp. 10684–10695.
[3] G. Pennycook and D. G. Rand, ‘‘The psychology of fake news,’’ *Trends Cogn. Sci.*, vol. 25, no. 5, pp. 388–402, May 2021.
[4] B. Singh and D. K. Sharma, ‘‘Predicting image credibility in fake news over social media using multi-modal approach,’’ *Neural Comput. Appl.*,
vol. 34, no. 24, pp. 21503–21517, Dec. 2022.
[5] N. Bonettini, P. Bestagini, S. Milani, and S. Tubaro, ‘‘On the use of Benford’s law to detect GAN-generated images,’’ in *Proc. 25th Int. Conf.*
Pattern Recognit. (ICPR), Jan. 2021, pp. 5495–5502.
[6] D. Deb, J. Zhang, and A. K. Jain, ‘‘AdvFaces: Adversarial face synthesis,’ in *Proc. IEEE Int. Joint Conf. Biometrics (IJCB)*, Sep. 2020, pp. 1–10.
[7] M. Khosravy, K. Nakamura, Y. Hirose, N. Nitta, and N. Babaguchi, ‘‘Model inversion attack: Analysis under gray-box scenario on deep
learning based face recognition system,’’ *KSII Trans. Internet Inf. Syst.*, vol. 15, no. 3, pp. 1100–1118, Mar. 2021.
[8] J. J. Bird, A. Naser, and A. Lotfi, ‘‘Writer-independent signature verification; evaluation of robotic and generative adversarial attacks,’’ *Inf.*
Sci., vol. 633, pp. 170–181, Jul. 2023.
[9] A. Ramesh, M. Pavlov, G. Goh, S. Gray, C. Voss, A. Radford, M. Chen, and I. Sutskever, ‘‘Zero-shot text-to-image generation,’’ in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 8821–8831.
[10] C. Saharia, W. Chan, S. Saxena, L. Li, J. Whang, E. Denton, S. K. S. Ghasemipour, B. K. Ayan, S. S. Mahdavi, R. G. Lopes,
T. Salimans, J. Ho, D. J. Fleet, and M. Norouzi, ‘‘Photorealistic textto- image diffusion models with deep language understanding,’’ 2022,
arXiv:2205.11487.
[11] P. Chambon, C. Bluethgen, C. P. Langlotz, and A. Chaudhari, ‘‘Adapting pretrained vision-language foundational models to medical imaging
domains,’’ 2022, *arXiv:2210.04133*.
[12] F. Schneider, O. Kamal, Z. Jin, and B. Schölkopf, ‘‘Moûsai: Text-to-music generation with long-context latent diffusion,’’ 2023, *arXiv:2301.11757*.
[13] F. Schneider, ‘‘ArchiSound: Audio generation with diffusion,’’ M.S. thesis, ETH Zurich, Zürich, Switzerland, 2023.
[14] D. Yi, C. Guo, and T. Bai, ‘‘Exploring painting synthesis with diffusion models,’’ in *Proc. IEEE 1st Int. Conf. Digit. Twins Parallel Intell. (DTPI)*, Jul. 2021, pp. 332–335.
[15] C. Guo, Y. Dou, T. Bai, X. Dai, C. Wang, and Y. Wen, ‘‘ArtVerse: A paradigm for parallel human–machine collaborative painting creation
in Metaverses,’’ *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 53, no. 4, pp. 2200–2208, Apr. 2023.
[16] Z. Sha, Z. Li, N. Yu, and Y. Zhang, ‘‘DE-FAKE: Detection and attribution of fake images generated by text-to-image generation models,’’ 2022,
arXiv:2210.06998.
[17] R. Corvi, D. Cozzolino, G. Zingarini, G. Poggi, K. Nagano, and L. Verdoliva, ‘‘On the detection of synthetic images generated by diffusion
models,’’ 2022, *arXiv:2211.00680*.
[18] I. Amerini, L. Galteri, R. Caldelli, and A. Del Bimbo, ‘‘Deepfake video detection through optical flow based CNN,’’ in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshop (ICCVW)*, Oct. 2019, pp. 1205–1207.
[19] D. Güera and E. J. Delp, ‘‘Deepfake video detection using recurrent neural networks,’’ in *Proc. 15th IEEE Int. Conf. Adv. Video Signal Based Surveill. (AVSS)*, Nov. 2018, pp. 1–6.
[20] J. Wang, Z. Wu, W. Ouyang, X. Han, J. Chen, Y.-G. Jiang, and S.-N. Li, ‘‘M2TR: Multi-modal multi-scale transformers for Deepfake detection,’’ in *Proc. Int. Conf. Multimedia Retr.*, Jun. 2022, pp. 615–623.
[21] P. Saikia, D. Dholaria, P. Yadav, V. Patel, and M. Roy, ‘‘A hybrid CNNLSTM model for video Deepfake detection by leveraging optical flow

features,'' in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2022, pp. 1–7.
[22] H. Li, B. Li, S. Tan, and J. Huang, ''Identification of deep network generated images using disparities in color components,'' *Signal Process.*,
vol. 174, Sep. 2020, Art. no. 107616.
[23] S. J. Nightingale, K. A. Wade, and D. G. Watson, ''Can people identify original and manipulated photos of real-world scenes?'' *Cognit. Res.*,
Princ. Implications, vol. 2, no. 1, pp. 1–21, Dec. 2017.
[24] A. Krizhevsky and G. Hinton, ''Learning multiple layers of features from tiny images,'' 2009.
[25] C. Schuhmann, R. Beaumont, R. Vencu, C. Gordon, R. Wightman, M. Cherti, T. Coombes, A. Katta, C. Mullis, M. Wortsman,
P. Schramowski, S. Kundurthy, K. Crowson, L. Schmidt, R. Kaczmarczyk, and J. Jitsev, ''LAION-5B: An open large-scale dataset for training next
generation image-text models,'' 2022, *arXiv:2210.08402*.
[26] Y. LeCun, Y. Bengio, and G. Hinton, ''Deep learning,'' *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.
[27] J. Gu, Z. Wang, J. Kuen, L. Ma, A. Shahroudy, B. Shuai, T. Liu, X. Wang,
G. Wang, J. Cai, and T. Chen, ''Recent advances in convolutional neural networks,'' *Pattern Recognit.*, vol. 77, pp. 354–377, May 2018.
[28] Z. Li, F. Liu, W. Yang, S. Peng, and J. Zhou, ''A survey of convolutional neural networks: Analysis, applications, and prospects,'' *IEEE Trans.*
Neural Netw. Learn. Syst., vol. 33, no. 12, pp. 6999–7019, Dec. 2022
[29] D. Gunning, M. Stefik, J. Choi, T. Miller, S. Stumpf, and G.-Z. Yang, ''XAI—Explainable artificial intelligence,'' *Sci. Robot.*, vol. 4, no. 37,
Dec. 2019, Art. no. eaay7120.
[30] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, ''Grad-CAM: Visual explanations from deep networks via
gradient-based localization,'' in *Proc. IEEE Int. Conf. Comput. Vis.*
(ICCV), Oct. 2017, pp. 618–626.
[31] M. Abadi et al. (2015). *TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems*. [Online]. Available: https://www.google.com/