IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received September 2, 2020, accepted September 10, 2020, date of publication September 18, 2020,
date of current version October 1, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3024575

Investigating and Suggesting the Evaluation Dataset for Image Classification Model

SARASWATHI SIVAMANI, SUN IL CHON, DO YEON CHOI, AND JI HWAN PARK
ThinkforBL Consultancy Services, Seoul 06236, South Korea

Corresponding author: Ji Hwan Park (jihwan.park@thinkforbl.com)

ABSTRACT Image processing systems are widespread with the digital transformation of artificial intelligence. Many researchers developed and tested several image classification models using machine learning and statistical techniques. Nevertheless, the current research seldom focuses on the quality assurance of these models. The existing methods lack to verify the quality assurance, with the lack of test cases to prepare the evaluation dataset to test the model, which can cause critical drawbacks in the nuclear field and defense system. In this article, we discuss and suggest the preparation of the evaluation dataset using improved test cases through Cause-Effect Graphing. The proposed method can generate the evaluation dataset with automated test cases through the quantification method, which consists of 1) image characteristic selection 2) creating the Cause-Effect graphing approach of the image with the feature, and 3) generate all possible test coverage. The testing is performed with the COCO dataset, which shows the declining prediction accuracy with the adjusted brightness and sharpness ranging between −75 to 75%, which indicates the negligence of the important characteristics in the existing test dataset. The experiment shows the prediction fails while sharpness is less than the 0%, and the brightness fails at −75% with less number of detection object between −50% and 75%. This indicates that characteristic changes affects the prediction accuracy and the number of detected objects in an image. Our approach proves the importance of the characteristic selection process for the overall image to generate a more efficient model and increase the accuracy of object detection.

INDEX TERMS Machine learning, quality assurance, cause-effect graphing, object detection, image classification.

I. INTRODUCTION

Advances in machine learning (ML) techniques have spread to the wide range of applications that perform advanced perception and decision-making in various important security-related fields [1]. Safety-critical machine learning systems needs best quality assurance methods for identifying the risks on real-time. Currently, the methods used to identify the model accuracy includes precision-recall, F-measure, and ROC curve [2]–[4]. Quality assurance (QA) has a high effect on social acceptance because it has always been a way to deliver safety and security [5].

Research on the dataset was mostly focused on the dataset balance, feature selection and skewness [6]. Among them, feature selection is very important process which identifies and removes the irrelevant features. Ghotra *et al.* [7] developed metrics to confirm the impact of the feature selection in the dataset, by comparing the accuracy with 30 feature selection techniques. Attaching the importance to the feature selection, many researchers are focused on new feature selection technique based on machine learning such as Random forest [8], Decision tree [9], MapReduce [10] and so on. Similarly, the scope of research mostly focuses on the balance between the training and testing dataset, explaining its importance in the test accuracy and image classification [11]. Ponce *et al.* [12] makes a detailed explanation of the dataset issues in the object recognition, especially with the image dataset such as Caltech and PASCAL. It mainly focuses on the object detection with different background. The issues in the Corel dataset and its annotations were analyzed using Support vector machine (SVM) [13]. Zhang *et al.* [14] also trained and tested the PASCAL dataset with different combination of the background. Catal and Diri [15] investigate the effect of dataset size and metrics for the fault prediction problem,

The associate editor coordinating the review of this manuscript and approving it for publication was Hossein Rahmani.

while Bennin *et al.* [16] investigate the effect of the balanced training with the dataset on the prediction system.

An improper data preparation can decrease accuracy and increase the number of errors [17]. Recently, researchers focus on the characteristics of the dataset. Oreski *et al.* [18] emphasis the importance of the dataset characteristics for the feature selection but ignore the image characteristic. Dodge and Karam [19] examines the importance of the image quality that affects the accuracy in deep neural network. Five types of image quality distortions such as blur, noise, contrast, JPEG and JPEG2000 compression are used on different deep learning techniques. Machine learning algorithms are also used to analyze and improvise the image quality [20] and image denoising [21]. All these studies reveal there is no single method to prepare the balanced dataset with required image characteristics. Hence, in this article, we focus on the impact of image characteristic to build a balanced dataset.

Most of the researchers claim the highest performance for their model, but some reveals unsuccessful result for the same images in a different environment [22]. Even though some of the researchers claimed that their models provided the highest performance, these models are proved to be unsuccessful when evaluated on public datasets [23]. Therefore, it is extremely crucial to benchmark available fault prediction models under different conditions on public datasets. We acknowledge the availability of effort-aware predictive models but we argue that the models selected to cover imbalance dataset with different domains of techniques ranging from statistical, data mining, and machine learning techniques. Hence these models become the representative of all the several models that exist in literature.

The current regulatory framework for different kinds of software relies on a software and system engineering paradigm that was clearly not designed with machine learning in mind. Widely used standards for software development life-cycle processes, like IEC62304 for medical device software [24], ISO26262 [25] for automotive, or ISO25000 [26] for general-purpose software, are based on defining requirements, defining the architecture, decomposing the system into smaller units, integrating, verifying and validating the results. From the 5V's [27] of Bigdata such as Variety, Velocity, Volume, Veracity, and Volume, Variety is the most important feature bound for quality assurance. Therefore, in this article, we focus on a variety of datasets.

In fact, recent accidents caused during the use of several experimental on security have revealed QA frameworks as imperative for addressing this upcoming social issue [28]. Although the image classification in security-related fields are actively developed and proposed, QA concept and technologies to ensure safety and security have not been systematized yet. Therefore, in this study, we organize the review for the open QA problems on safety-critical machine learning systems using public image dataset COCO with machine-learning models as an example. Kim [29] has validated the COCO dataset using the YOLO model, and explained the labelling issues. However, the COCO dataset has been used in many image classification analysis [30]–[36]. Pont-Tuset and Van Gool [37] also boost the preference of COCO from PASCAL. We also chose the COCO dataset to our analysis. The contributions of this study are:

- Clarification of the problems related to the quality assurance of the trained model with the given dataset in image classification.
- Problems on the safety-related system, especially with machine learning models.
- Shown the effectiveness of Cause-effect graphing testing to prepare the evaluation dataset;
- Shown the open issues in the current testing and next research direction.

II. RESEARCH QUESTIONS

Most research presumes that the prediction models will be improved, if the feature selection of appropriate classes for the dataset are selected and trained, high performance can be provided. We embark on experimenting with such a model to prove the importance of the evaluation dataset with a different characteristic of the image in the overall dataset. The objective of the paper is to explain that the preparation of the evaluation dataset is equally important with the preparation of the training dataset. The objective was inspired by the following questions, which will be discussed later in the result.

1. Can the trained model detect all the objects in the image, if there is a change in the image characteristics? In case, the image was taken in a night or darker area?

2. Could the possible characteristics of the image be identified to generate an accurate model? The search for characteristics is not limited and can even be considered as a feature selection of the dataset. Initially, it may be reasonable to think that the dataset contains all the possible characteristics that would be identified to produce the best model, but that might not be always the case.

3. What if the model failed in the crucial industry connected with national security and health care centers? The image classification can be sensitive in such a field, which needs more distinct and high importance to the models. But, it can be seen that most of the public dataset does not produce high accuracy when tested with different images from different places. First, it is important to consider all possible characteristics of overall images together with the feature selection of the dataset.

4. How to obtain more accurate results? It may seem justified that the training dataset should contain all the characteristic of the images, to produce an efficient model. We explore this domain, to investigate the evaluation dataset on the prediction model for high accuracy result.

III. EXPERIMENTAL SETTING

First, we describe the datasets used in the study. We conducted the experiment with the open access COCO dataset. The evaluation dataset was prepared with Cause-Effect

Graphing. YOLO framework was used to train and test the model.

A. DATA COLLECTION

The Common Objects in Context dataset (COCO) [38] basically has 91 object classes that have more than 5,000 labeled instances. From the total image of 328,000 images, there are 2,500,000 labeled instances. COCO dataset consists of fewer categories compared to the ImageNet dataset [39] but rich in instances, which improves learning object models for precision and accuracy. The dataset is also larger in the number of instances compared with the PASCAL VOC [40] and SUN [41] datasets. In addition, the COCO dataset can be used for learning contextual information, which has considerably more object instances per image than ImageNet and PASCAL. Although the SUN dataset contains significant contextual information, very few instances. Therefore, we chose the COCO dataset in our study.

B. CAUSE-EFFECT GRAPHING

Cause-Effect Graphing [42] graphically represent the situations of combinations of various input conditions and the issues that manipulate the result. The graph is later transformed into a decision table to obtain the number of test cases. The cause-effect graphing technique is used because boundary value analysis and equivalence class partitioning methods do not consider the combinations of input conditions. But since there may be some critical behavior to be tested when some combinations of input conditions are considered, that is why the cause-effect graphing technique is used.

Steps used in deriving test cases using this technique are:

1. Divides the specifications into small pieces and change them into the cause-effect graphs.
2. Identify the possible input and effective output conditions in the specifications.
3. With Boolean expressions, the specifications are transformed into a cause-effect graph, where the constraint is added at the required places.
4. Convert the graph into a decision table.
5. Conclude the decision table into potential test cases

Although the method can detect the ambiguity and inadequacy with a unique perspective, it cannot guarantee that all the test cases are useful.

C. CHARACTERISTIC ANALYSIS

Identifying the characteristic of the image is similar to the feature selection of the images. Feature selection is the process of selecting the most important feature that will have an impact on the model performance such as overfitting, accuracy, and mainly manage the training time. Similarly, the characteristic of every image is important which needs to identified and included in the training and evaluation dataset, which will greatly affect the accuracy of the prediction model.

When the COCO dataset is used in our experiment, first we analyze and evaluate the characteristic of the image. Without a doubt the characteristic of the objects is also more important for accuracy, which will be analyzed and discussed in our future work with other comparative datasets. From 91 objects in the COCO, we chose room images for our analysis, which contains a tv monitor, dining table, chair, vase, clock, refrigerator, potted plant, and people. With meticulous observation and analysis, we concluded the characteristic of the images into three causes such as brightness, sharpness, and grayscale. The brightness and sharpness of the images are altered on different range and verified with the predefined test cases scenario through cause-effect graphing.

IV. EXPERIMENT DESCRIPTION

The experiment revolves around the analysis of the prediction accuracy of the model with different scenarios involving the characteristic difference of the model. The YOLO –Darknet framework [43] was used to test and compare the model with the normal image and the new images with adjusted brightness and sharpness of the image. In this case, more types of datasets are required to improve the accuracy of object detection.

The table 1 shows the confidence score of the object detected in the brightness adjusted image, which was obtained from the COCO dataset. Among the many characteristics, we have chosen two causes brightness and sharpness to test and compare the test images. The same result is shown in table 2 with the detected image, along with the bounding boxes. From the result, we can clearly see the reduction of detected objects, with respect to the changes in the brightness of the images. Also, the confidence score is declining on each adjusted image. According to the result, it can be concluded that the object detection is greatly affected below and above 50%. Although there are few object detections, the prediction percentage is affected.

Similarly, the sharpness of the image is adjusted between −75 and 75% and the results are tabulated in table 3 and 4. Contradicting to the brightness, the sharpness adjusted images are more stable with increment, but decrement of changes in sharpness percentage exhibits failure in detection. This implies that the training dataset lacks images with different characteristic, which may affect the training model and test results.

Testing a machine learning model has several potential limitations. Therefore, we discuss the possible threats to identify the object in the images. This study identifies the faults in dataset preparation through the comparison of brightness/sharpness corrected images with the original images. In this case, more types of datasets are required to improve the accuracy of object detection

Considering the original images, it is difficult to generalize our accuracy of the trained model. We need to prepare an evaluation dataset that can test all possible characteristic of the images. For this reason, we use the CETA (Cause-Effect Test Analysis) tool [44], drafting all possible test cases to verify the accuracy of the object detection.

TABLE 1. Confidence level of each object in an image with different brightness levels.

Results \ Brightness		-75	-50	-25	0	25	50	75
Test Case[1]		117-13	3-4					117-14
		Fail	Pass	Pass	Pass	Pass	Pass	Pass
Object Detection[2]	Bottle	-	58	60	-	-	58	69
	Vase1	-	91	93	84	85	69	-
	Vase2	-	80	70	61	59	-	-
	Vase3	-	-	-	59	-	-	-
	Vase4	-	-	-	52	-	-	-
	Clock	-	-	60	73	54	66	-
	Refrigerator	-	-	-	76	-	-	-
	Tvmonitor1	-	93	98	99	99	98	96
	Tvmonitor2	-	-	52	55	-	-	-
	Diningtable1	-	-	67	70	76	70	56
	Diningtable2	-	-	67	57	-	-	-
	Pottedplant	-	-	60	63	77	75	-
	Chair1	-	97	100	100	100	99	100
	Chair2	-	96	97	98	96	97	94
	Chair3	-	96	97	97	95	96	88
	Person1	-	85	94	95	96	94	-
	Person2	-	82	92	91	85	-	-

[1] Test case result which derived from Cause and Graphing test case design method
[2] Object detected confidence using YOLO algorithm

Using the CETA tool, which was developed from our earlier research, we have derived all possible test cases, including the characteristics of the images and objects. This quality assurance tool creates test cases with the help of cause-effect graphing. Except the time to identify and input the characteristic in the tool, the speed of the test case generation is very fast with negligible time complexity, owing to the graphing method. Figure 1 shows the test case generation in the tool with 13 causes and 2 effects. The two effects represent the result of whether the objects in the image are detected or not detected.

The layout of the tools has 6 windows that describes

1) Characteristics of Cause-effect for the images: It gives the detailed description of the specification and conditions of cause-effect.
2) Contraction of the Cause-effect contains the cause and effect list.
3) Test Scenario
4) Test Values
5) Cause-Effect Graph
6) Test Cases.

By adjusting the brightness and sharpness of the image between −75 to 75%, the test cases are derived for the COCO dataset using the cause-effect method and tabulated in the table 5. The characteristics are categorized as the causes ranging from C1 to C13, in which the effects are divided into E1 and E2. The total test cases obtained range to 6,345 with 140 test scenario generated from decision table that consist of 13 causes and 2 effects. When the number of objects detects exceeds 50%, with the combination of all image characteristics, the expected result of test case succeeds. With these test cases, the evaluation dataset can be prepared to test the model for the prediction accuracy.

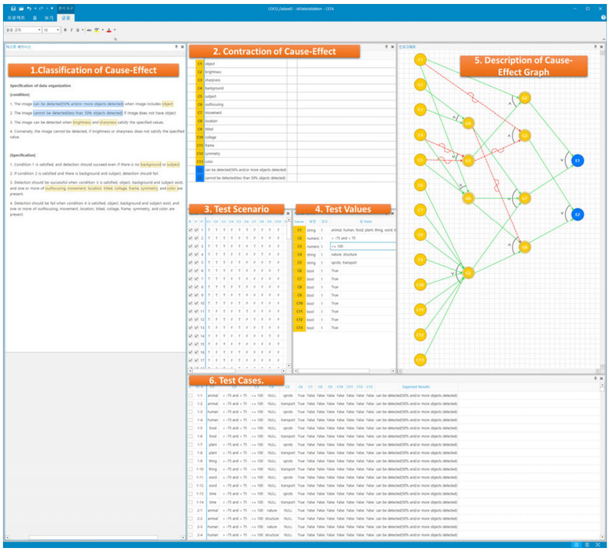

1.Classification of Cause-Effect
2. Contraction of Cause-Effect
3. Test Scenario
4. Test Values
5. Description of Cause-Effect Graph
6. Test Cases.

FIGURE 1. Overview of the Cause-effect graph in CETA tool.

TABLE 2. Test result of detection objects in an image at different level of brightness.

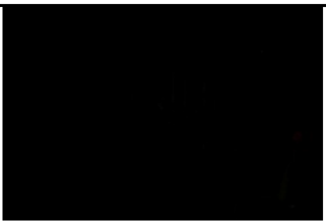
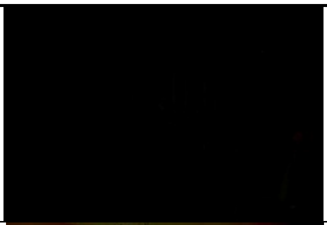
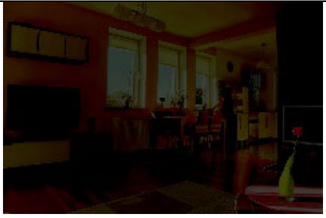
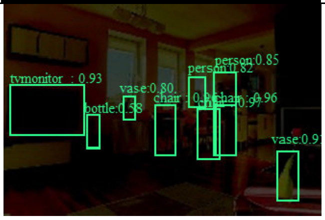
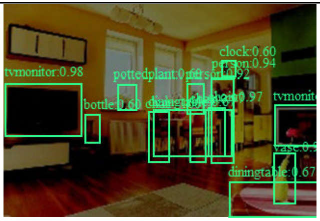
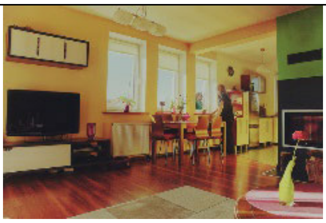
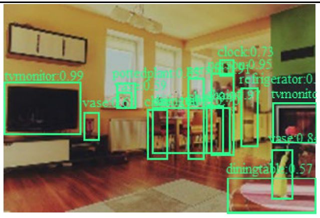
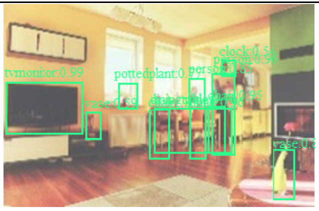
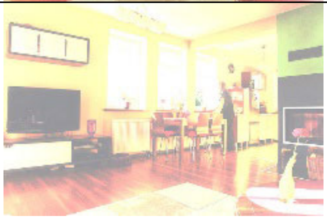
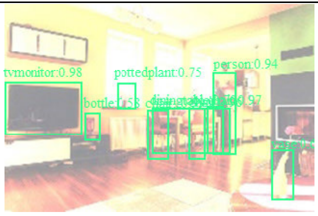
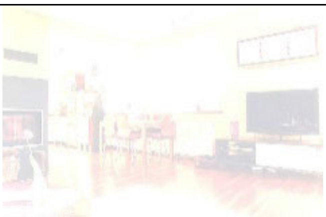
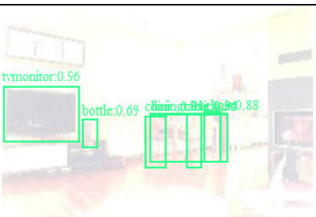
Brightness (%)	Brightness adjusted test images	Detected images	No of detected objects
-75			0
-50		 tvmonitor : 0.93 bottle:0.58 vase:0.80 person:0.85 person:0.82 chair:0.96 chair:0.97 vase:0.9	9
-25		 clock:0.60 person:0.94 tvmonitor:0.98 person:0.92 bottle:0.61 chair:0.97 tvmonit diningtable:0.67	14
0 (original)		 clock:0.73 tvmonitor:0.99 refrigerator:0. tvmonito vase:0.8 diningtable:0.57	16
25			11
50		 tvmonitor:0.98 pottedplant:0.75 person:0.94 bottle:0.58	11
75			6

TABLE 3. Confidence level of each object in an image with different Sharpness levels.

Sharpness Results		-75	-50	-25	0	25	50	75
Test Case		127-14	3-1					127-13
		Fail	Fail	Fail	Pass	Pass	Pass	Pass
Object Detection	Bottle	-	-	-	-	74	82	91
	Vase1	-	-	-	84	75	75	65
	Vase2	-	-	-	61	71	69	64
	Vase3	-	-	-	59	-	-	-
	Vase4	-	-	-	52	-	-	-
	Clock	-	-	-	73	70	62	-
	Refrigerator	-	-	-	76	63	-	-
	Tvmonitor1	-	-	-	99	100	100	100
	Tvmonitor2	-	-	-	55	63	-	-
	Diningtable1	-	-	-	70	66	67	58
	Diningtable2	-	-	-	57	63	61	58
	Pottedplant	-	-	-	63	64	59	56
	Chair1	-	-	-	100	100	100	99
	Chair2	-	-	-	98	98	98	98
	Chair3	-	-	-	97	98	97	97
	Person1	-	-	-	95	98	97	96
	Person2	-	-	-	91	92	93	88

V. ANALYSIS OF EXPERIMENT

Comparison performance of the evaluation dataset with and without the test cases has given detailed knowledge and importance towards the preparation of the dataset. Research questions that motivated the study are explained in detail with the experimental result.

Question 1: *Does the training dataset include all possible characteristics?*

We compared the result of the original image and the altered images in the same trained YOLO model as shown in Tables 2 and 4. The result shows that the confidence score changes with minimal changes of brightness and sharpness in the images, while the deduction fails with the maximum changes. From the table, it is evident that the training dataset does not include all possible characteristics in the dataset, resulting in the low accuracy of detection results.

Question 2*. Could the possible characteristics of the image be identified to generate an accurate model?*

The pre-trained model in the YOLO consists of 91 categories of images from different sources. For object detection, the model is undoubtedly a well-trained model that can detect most of the objects with the better accuracy. However, examining the results of the comparison table 1, it could be observed that the performance of the detection is inconsistent. In other words, we need to include more training data with the additional characteristics of the images.

Question 3: *What if the model failed in the crucial industry connected with national security and health care centers?*

Pattern recognition has taken a significant role in the many high platforms such as national security, nuclear energy, and the medical field. With advanced methods and algorithms, the trust of the machine learning algorithm has accepted. Hence, if a simple characteristic is missed during the training, it may cause a major issue in the high-risk field. To emphasis the importance of the issue, the paper compares the result of the altered image with the actual image. With the Cause-effect graphing tool, it becomes even more efficient to assure the quality of the trained model.

Question 4: *How to obtain more accurate results?*

To find out whether the model can produce more accurate results, we possibly need a quality assurance method. Quality assurance for the machine learning technique has always reminded a challenge for the following reason. If the training dataset does not include all the features of the images, then it will not produce an accurate result.

To minimize the probability of failures, ML models has to be tested with balanced dataset, especially for sensitive domains. Generally speaking, the outcome of an ML model is a prediction, which is not easy to compare or verify against some kind of expected value. Nevertheless, developers test the machine learning model performance by comparing predicted values with the model output values, which is different from testing the ML model for any input, due to its limitation. The so-called black box testing of ML models can employ a variety of techniques, such as metamorphic testing, model performance, dual coding, comparison with linear models, and coverage guided fuzzing and testing with varying data slices. There is also the problem of causality. A machine learning algorithm doesn't know if a regularity found on input data is a cause for a prediction or just

TABLE 4. Test result of detection objects in an image at different level of sharpness.

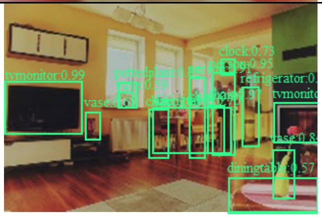
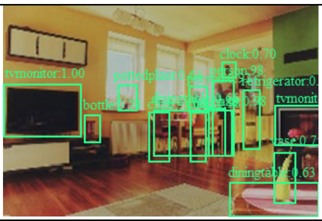
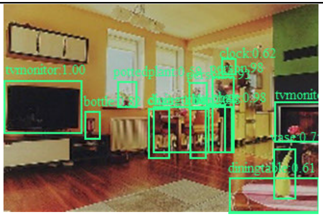
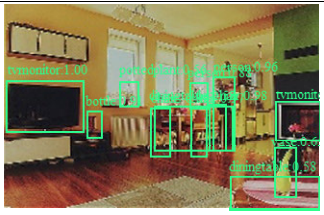
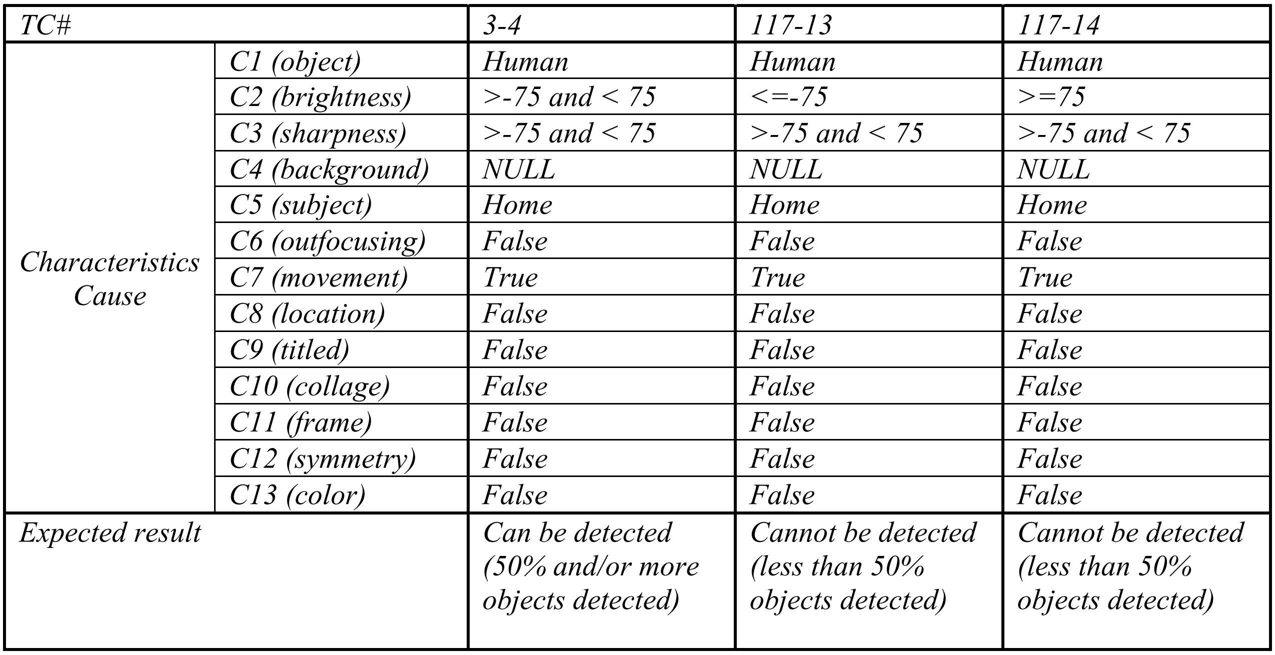
Sharpness(%)	Sharpness adjusted test images	Detected images	No of detected objects
-75			0
-50			0
-25			0
0 (original)		 tvmonitor:0.99 clock:0.73 person:0.95 refrigerator:0. vase:0.59 tvmonito vase:0.8 diningtable:0.57	16
25		 tvmonitor:1.00 clock:0.70 person:98 pottedplant:0.64 refrigerator:0. bottle:0.74 tvmonit vase:0.7 diningtable:0.63	15
50		 tvmonitor:1.00 clock:0.62 person:98 bottle:0.82 tvmonite vase:0.7 diningtable:0.61	13
75		 tvmonitor:1.00 person:0.96 tvmonite vase:0.6 diningtable:0.58	12

TABLE 5. Target test case for the COCO dataset.



TC#		*3-4*	*117-13*	*117-14*
Characteristics Cause	*C1 (object)*	*Human*	*Human*	*Human*
	C2 (brightness)	*>-75 and < 75*	*<=-75*	*>=75*
	C3 (sharpness)	*>-75 and < 75*	*>-75 and < 75*	*>-75 and < 75*
	C4 (background)	*NULL*	*NULL*	*NULL*
	C5 (subject)	*Home*	*Home*	*Home*
	C6 (outfocusing)	*False*	*False*	*False*
	C7 (movement)	*True*	*True*	*True*
	C8 (location)	*False*	*False*	*False*
	C9 (titled)	*False*	*False*	*False*
	C10 (collage)	*False*	*False*	*False*
	C11 (frame)	*False*	*False*	*False*
	C12 (symmetry)	*False*	*False*	*False*
	C13 (color)	*False*	*False*	*False*
Expected result		*Can be detected (50% and/or more objects detected)*	*Cannot be detected (less than 50% objects detected)*	*Cannot be detected (less than 50% objects detected)*

a correlation. Thus, making the quality assurance difficult for the ML-based models.

In accordance with this issue, we use the cause-effect graphing tool to create the test cases to prepare the evaluation dataset and test all possible outcomes of the images with the public COCO dataset. The comparison result shows that the need for characteristic identification and the cause-effect graphing tools help to create the evaluation dataset to test the results. The existing research focus on the quality and accuracy for the machine learning models in literature but our study explains that the missing characteristic of the images can disturb the accuracy of the object detection and cause an imbalanced learning. We intend to consider more unique characteristics in various datasets in future studies.

VI. CONCLUSION

We have analyzed and examined the importance of preparing the evaluation dataset through the cause-effect testing method. COCO open dataset was used for the experiment to test two types of scenarios indulging the characteristics of the image. To our interest were how the percentage of the test dataset differed with the prediction accuracy in the images. The two characteristic such as brightness and sharpness were identified to respond to the research question, while CETA tool was used to create the test cases with the various characteristics. The result shows the decline in the prediction accuracy with respect to the changes in the adjusted images, indicating the importance of the image characteristics. To the best of our knowledge, this is the study that focuses on the preparation of evaluation dataset with test cases that includes the characteristics of the images. Assessing the performance of the testing with and without the adjusted images, our result derived to a conclusion that the dataset can achieve a better result with a well-prepared train dataset and evaluation dataset, using the cause-effect based test cases.

REFERENCES

[1] S. Saran, N. Natarajan, and M. Srikumar, "In pursuit of autonomy: AI and national strategies," Observer Res. Found., New Delhi, India, ORF Special Rep. 76, Nov. 2018.

[2] D. M. Powers, "Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation," *J. Mach. Learn. Technol.*, vol. 2, pp. 37–63 2011.

[3] C. Goutte and E. Gaussier, "A probabilistic interpretation of precision, recall and F-score, with implication for evaluation," in *Proc. Eur. Conf. Inf. Retr.* Berlin, Germany: Springer, Mar. 2015, pp. 345–359.

[4] T. Saito and M. Rehmsmeier, "Precrec: Fast and accurate precision–recall and ROC curve calculations in R," *Bioinformatics*, vol. 33, no. 1, pp. 145–147, 2017.

[5] L. Ma, F. Juefei-Xu, M. Xue, Q. Hu, S. Chen, B. Li, Y. Liu, J. Zhao, J. Yin, and S. See, "Secure deep learning engineering: A software quality assurance perspective," 2018, *arXiv:1810.04538*. [Online]. Available: http://arxiv.org/abs/1810.04538

[6] N. E. Fenton and N. Ohlsson, "Quantitative analysis of faults and failures in a complex software system," *IEEE Trans. Softw. Eng.*, vol. 26, no. 8, pp. 797–814, Aug. 2000.

[7] B. Ghotra, S. McIntosh, and A. E. Hassan, "A large-scale study of the impact of feature selection techniques on defect classification models," in *Proc. IEEE/ACM 14th Int. Conf. Mining Softw. Repositories (MSR)*, May 2017, pp. 146–157.

[8] M. A. M. Hasan, M. Nasser, S. Ahmad, and K. I. Molla, "Feature selection for intrusion detection using random forest," *J. Inf. Secur.*, vol. 7, no. 3, pp. 129–140, 2016.

[9] H. Rao, X. Shi, A. K. Rodrigue, J. Feng, Y. Xia, M. Elhoseny, X. Yuan, and L. Gu, "Feature selection based on artificial bee colony and gradient boosting decision tree," *Appl. Soft Comput.*, vol. 74, pp. 634–642, Jan. 2019.

[10] C. Reggiani, Y. A. Le Borgne, and G. Bontempi, "Feature selection in high-dimensional dataset using MapReduce," in *Proc. Benelux Conf. Artif. Intell.* Cham, Switzerland: Springer, 2017, pp. 101–115.

[11] D. Kang and S. Oh, "Balanced training/test set sampling for proper evaluation of classification models," *Intell. Data Anal.*, vol. 24, no. 1, pp. 5–18, 2020.

[12] J. Ponce, T. L. Berg, M. Everingham, D. A. Forsyth, M. Hebert, S. Lazebnik, M. Marszalek, C. Schmid, B. C. Russell, A. Torralba, C. K. I. Williams, J. Zhang, and A. Zisserman, "Dataset issues in object recognition," in *Toward Category-Level Object Recognition*. Berlin, Germany: Springer, 2006, pp. 29–48.

[13] J. Tang and P. H. Lewis, "A study of quality issues for image auto-annotation with the corel dataset," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 17, no. 3, pp. 384–389, Mar. 2007.

[14] J. Zhang, M. Marszałek, S. Lazebnik, and C. Schmid, "Local features and kernels for classification of texture and object categories: A comprehensive study," *Int. J. Comput. Vis.*, vol. 73, no. 2, pp. 213–238, Jun. 2007.

[15] C. Catal and B. Diri, "Investigating the effect of dataset size, metrics sets, and feature selection techniques on software fault prediction problem," *Inf. Sci.*, vol. 179, no. 8, pp. 1040–1058, Mar. 2009.

[16] K. E. Bennin, J. Keung, A. Monden, Y. Kamei, and N. Ubayashi, "Investigating the effects of balanced training and testing datasets on effort-aware fault prediction models," in *Proc. IEEE 40th Annu. Comput. Softw. Appl. Conf. (COMPSAC)*, vol. 1, Jun. 2016, pp. 154–163.

[17] C. R. Corbeil, C. I. Williams, and P. Labute, "Variability in docking success rates due to dataset preparation," *J. Comput.-Aided Mol. Des.*, vol. 26, no. 6, pp. 775–786, Jun. 2012.

[18] D. Oreski, S. Oreski, and B. Klicek, "Effects of dataset characteristics on the performance of feature selection techniques," *Appl. Soft Comput.*, vol. 52, pp. 109–119, Mar. 2017.

[19] S. Dodge and L. Karam, "Understanding how image quality affects deep neural networks," in *Proc. 8th Int. Conf. Quality Multimedia Exp. (QoMEX)*, Jun. 2016, pp. 1–6.

[20] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, "Analyzing and improving the image quality of StyleGAN," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2020, pp. 8110–8119.

[21] A. Abdelhamed, M. Afifi, R. Timofte, and M. S. Brown, "NTIRE 2020 challenge on real image denoising: Dataset, methods and results," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops*, Jun. 2020, pp. 496–497.

[22] M. Nilashi, O. Ibrahim, M. Dalvi, H. Ahmadi, and L. Shahmoradi, "Accuracy improvement for diabetes disease classification: A case on a public medical dataset," *Fuzzy Inf. Eng.*, vol. 9, no. 3, pp. 345–357, Sep. 2017.

[23] C. Kolias, G. Kambourakis, A. Stavrou, and S. Gritzalis, "Intrusion detection in 802.11 networks: Empirical evaluation of threats and a public dataset," *IEEE Commun. Surveys Tuts.*, vol. 18, no. 1, pp. 184–208, 1st Quart., 2016.

[24] P. Jordan, "Standard IEC 62304—Medical device software—Software lifecycle processes," Tech. Rep., 2006, pp. 41–47.

[25] R. Salay, R. Queiroz, and K. Czarnecki, "An analysis of ISO 26262: Using machine learning safely in automotive software," 2017, *arXiv:1709.02435*. [Online]. Available: http://arxiv.org/abs/1709.02435

[26] D. Zubrow, "Software quality requirements and evaluation, the ISO 25000 series," *Softw. Eng.*, 2004.

[27] Y. Demchenko, C. de Laat, and P. Membrey, "Defining architecture components of the big data ecosystem," in *Proc. Int. Conf. Collaboration Technol. Syst. (CTS)*, May 2014, pp. 104–112.

[28] D. Firesmith, "Engineering safety-and security-related requirements for software-intensive systems," Carnegie-Mellon Univ. Softw. Eng. Inst, Pittsburgh, PA, USA, Tech. Rep., 2007, p. 126.

[29] D. H. Kim, "Evaluation of COCO validation 2017 dataset with YOLOv3," *Evaluation*, vol. 6, no. 7, pp. 10356–10360, 2019.

[30] S. Shao, Z. Li, T. Zhang, C. Peng, G. Yu, X. Zhang, J. Li, and J. Sun, "Objects365: A large-scale, high-quality dataset for object detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2019, pp. 8430–8439.

[31] S. Rostianingsih, A. Setiawan, and C. I. Halim, "COCO (creating common object in context) dataset for chemistry apparatus," *Procedia Comput. Sci.*, vol. 171, pp. 2445–2452, Jan. 2020.

[32] Y. Yao, Y. Wang, Y. Guo, J. Lin, H. Qin, and J. Yan, "Cross-dataset training for class increasing object detection," 2020, *arXiv:2001.04621*. [Online]. Available: http://arxiv.org/abs/2001.04621

[33] P. Sharma, N. Ding, S. Goodman, and R. Soricut, "Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning," in *Proc. 56th Annu. Meeting Assoc. Comput. Linguistics*, vol. 1, 2018, pp. 2556–2565.

[34] K. Sun, B. Xiao, D. Liu, and J. Wang, "Deep high-resolution representation learning for human pose estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2019, pp. 5693–5703.

[35] R. Hu, P. Dollár, K. He, T. Darrell, and R. Girshick, "Learning to segment everything," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4233–4241.

[36] P. Sun, H. Kretzschmar, X. Dotiwalla, A, Chouard, V. Patnaik, P. Tsui, and J. Guo, "Scalability in perception for autonomous driving: Waymo open dataset," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2020, pp. 2446–2454.

[37] J. Pont-Tuset and L. V. Gool, "Boosting object proposals: From Pascal to COCO," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 1546–1554.

[38] A. Veit, T. Matera, L. Neumann, J. Matas, and S. Belongie, "COCO-text: Dataset and benchmark for text detection and recognition in natural images," 2016, *arXiv:1601.07140*. [Online]. Available: http://arxiv.org/abs/1601.07140

[39] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.

[40] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The Pascal visual object classes (VOC) challenge," *Int. J. Comput. Vis.*, vol. 88, no. 2, pp. 303–338, Jun. 2010.

[41] J. Xiao, J. Hays, K. A. Ehinger, A. Oliva, and A. Torralba, "SUN database: Large-scale scene recognition from abbey to zoo," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Jun. 2010, pp. 3485–3492.

[42] K. Nursimulu and R. L. Probert, "Cause-effect graphing analysis and validation of requirements," in *Proc. Conf. Centre Adv. Stud. Collaborative Res.*, Nov. 1995, p. 46.

[43] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," 2018, *arXiv:1804.02767*. [Online]. Available: http://arxiv.org/abs/1804.02767

[44] S. Chon and J. Park, "Suggestion of practical quantification measuring method of test design which can represent the current status," in *Proc. IEEE Int. Conf. Softw. Test., Verification Validation Workshops (ICSTW)*, Mar. 2017, pp. 294–299.

SARASWATHI SIVAMANI received the bachelor's degree in information technology from the Velammal Engineering College, India, in 2008, and the master's and Ph.D. degrees in information and communication engineering from Sunchon National University, South Korea, in 2015 and 2018, respectively. She has worked as Senior Engineer and a Java Developer for three years, from 2008 to 2011, at Tata Consultancy Services, India. Since 2018, she has been working as a Senior Researcher at ThinkforBL Consultancy Services. Her research interests include data analysis and image processing related to livestock behavior and welfare. Her knowledge also includes the ontology model, ubiquitous computing, and big data processing.

SUN IL CHON received the master's degree in electronics engineering from Jeonbuk National University. He is currently a Key Researcher with ThinkforBL Consultancy Services. He conducted consulting in SW engineering. He is researching the development of a smart barn service incorporating artificial intelligence SW technology.

DO YEON CHOI majored in software engineering at Jeonbuk National University and conducted consulting in software engineering and testing. She is currently working in a smart barn research incorporating AI technology and conducting quality assurance and software engineering activities, to lead a core researching in ThinkforBL Consultancy Services.

JI HWAN PARK majored in electronic at Sungkyunkwan University and conducted consulting in the SW engineering field of about 300 companies. He was an Adjunct Professor with the Department of Knowledge and Information Engineering, Ajou University, and the Department of Software Engineering, Jeonbuk National University. He is working on the application of SW technology to the field of smart livestock. He is currently the Vice President of the Korea Software Engineering Network (K.SEN) and the Korean Representative of the Aisa Software Quality Network (ASQN).