Exercise 8: Image Classification using Custom Images with K-Fold

Jasmin Clores   
College of Computer Science and Engineering  
Jose Rizal UniversityMandaluyong, Philippines  
jasmin.clores@my.jru.edu

Mark Joshua Lodriga  
College of Computer Science and Engineering  
Jose Rizal UniversityMandaluyong City, Philippines  
mark.lodriga@my.jru.edu

Mhargielyn Miñeque  
College of Computer Science and Engineering  
Jose Rizal UniversityMandaluyong City, Philippines  
mhargielyn.mineque@my.jru.edu

***Abstract*— This study examines the domain of image classification, specifically emphasizing the use of custom images through the application of k-fold cross-validation. It presents an innovative approach to building customized datasets, demonstrating the importance of modifying particular image classification tasks. This research also investigates various reviewed related literature wherein it has implemented or used a custom dataset within their specific domain and further analyzes how these studies can contribute to the overall field of this study. The methodology section includes essential experimental aspects such as preprocessing methods and architecture to obtain the best-performing model. Furthermore, this study explores the examination of two significant image classification techniques, which are the k-fold cross-validation and the holdout method. A thorough experimentation is performed, followed by the fine-tuning of hyperparameters and an in-depth analysis of the results for the top-performing model obtained. Overall, the study illustrates the complex aspects that define the model's performance under the two distinct techniques, and it aims to contribute valuable insights to the field of image classification by establishing the comparative strengths and limitations of k-fold cross-validation and holdout methods, ultimately guiding researchers in determining the most effective approach for a specific domain.**

**Keywords— Image Classification, Convolutional Neural Networks, Hold-out, K-fold (Cross-validation), Customized Dataset**

I. INTRODUCTION

The building of customized datasets has an integral part in the development and execution of accurate predictive models. This helps to organize data collection developed for different categories [1]. Developing a tailored dataset in the domain of machine learning is a detailed process of organizing various sets of data to meet the needs of a specific objective. By using this method for generating image datasets it can make a significant difference in the advancement of the ML model [2]. We utilize a prebuilt dataset of Garbage Images in relation to our chosen domain, which is easily accessible for this study. It provides several advantages in terms of improved efficiency of a model. Through the use of the chosen dataset, we aim to enhance the efficiency of implementing our ML models. This can be achieved by the dataset's prebuilt structure and labeling, which aligns with our study within our selected domain.

Creating a customized dataset for ML involves a complex process of different challenges. One of the primary challenges encountered in this field is the scale of a dataset that requires interaction with many other features to simulate applications in the real world [3]. In contrast to smaller datasets, which can be managed manually with a certain ease, managing huge data sets requires a thorough understanding of different features. Another problem that can be encountered is to the fundamental risk of bias. The potential for bias occurs in data collection and annotation, wherein subjective decisions can impact the dataset's structure [4]. Therefore, the biases in the training data can lead to a misunderstanding of the model's predictions or equality, primarily if the data represents certain cultural biases.

The hold-out method requires splitting the dataset among two distinct subsets. Training data is employed to acquire knowledge about the model, whereas testing data is collected for assessment purposes. On the other hand, k-fold involves the random partitioning of the dataset in 'k' subsets [5]. The motivation for detailed performance testing of the two validation methods comes from the complex way of achieving a balanced approach in model training. The key to success is achieving an ideal balance throughout the model's training and bypassing the challenge of overfitting by having an adequate number of instances. It also ensures accurate categorization when dealing with datasets of different sizes [6]. This emphasizes the need to evaluate the effectiveness of the hold-out and the K-Fold methods, which provides insight into their varying relevance for different dataset sizes.

II. LITERATURE REVIEW

A. Search for existing literature on building custom image datasets (minimum of 3 related papers to be reviewed, more will get additional points).

B. Summarize and Discuss the Highlights of the related papers

**Custom Dataset Creation with Tensorflow Framework and Image Processing for Google T-Rex**

This article examines a problem encountered by machine learning (ML) experts involved in the development of computer vision methods, which is specific to the lack of specialized datasets available for Google's T-Rex game [7]. In contrast to numerous image processing tasks that gain advantages from openly accessible datasets, this game lacks readily available data that complies with the unique input requirements for training deep neural networks. To address this existing research gap, the present study introduces a systematic strategy for developing a customized dataset for the T-Rex game. Thus, this technique provides a thorough overview of the systematic process for developing, training and evaluating a dataset using TensorFlow and Keras in Python. As a result, the method ensures integration across various computational architectures. This new strategy emphasizes the versatility of Convolutional Neural Networks (CNN) models in unexpected situations that showcase the ability of ML approaches to be customized for addressing difficulties in different applications when specific datasets are not commonly available.

**Construction of Diverse Image Datasets From Web Collections With Limited Labeling**

This research discusses some of the challenges of constructing detailed image datasets, which are important for developing image processing and multimedia studies. The conventional process of manually creating datasets is time-consuming and difficult to scale up [8]. As a result, there has been a rising demand for studying automated and semi-automatically methods using web images. However, the existing methodologies frequently generate datasets overloaded with unnecessary images or lack the important semantic and visual variety required to develop an adaptable classification.

In order to address these challenges, the study presents an innovative method that utilizes a semi-supervised coding structure. This system uses a uniform visual-semantic, which integrates visual features extracted from images and corresponding textual data obtained from online sources to manage a wide range of image datasets. The experimental findings show the system's effectiveness in creating and improving datasets with limited manual labeling. These results also indicate that the system performs better in terms of cross-dataset generalization and variety when compared to frequently utilized datasets in the real world.

**SODA: A large-scale open site object detection dataset for deep learning in construction**

This paper examines deep learning techniques utilized for product identification in construction. Wherein it presents a customized dataset named SODA, specifically tailored for huge-scale open-access image recognition. This dataset has 19,846 images and includes annotations for 286,201 objects categorized into 15 distinct categories relevant to building locations [9]. The development of this dataset was driven by the difficulty and security concerns associated with obtaining photos at building sites. This dataset addresses the drawbacks of previous datasets, characterized by their limited scope and range of categories. The study defines the systematic procedures entailed in selecting categories, acquiring data, cleaning and annotating the data, analyzing the information, and conducting experiments. Furthermore, the efficacy of the obtained dataset is evaluated through widely recognized object detection algorithms, showcasing consistent performance in the method selection process. Overall, this customed dataset is perceived as a valuable addition to the construction community, which aims to constantly enhance the range of construction-related objects, thereby promoting progress in the field of architecture.

**PadChest: A large chest x-ray image dataset with multi-label annotated reports**

This study presents PadChest, which has a detailed dataset of chest X-ray images designed to facilitate the automation analysis of healthcare visions and related information. The dataset consists of more than 160,000 pictures obtained from 67,000 patients.These images cover six diverse topics and are structured hierarchically [10]. Also, the customized dataset seeks to solve the limited availability of labeled public databases for chest X-ray images and equips a detailed annotation of 299 different medical entities. In addition, PadChest has many thoracic conditions and provides anatomical localization, high-resolution images, and patient information. It also represents the dataset's process that involves manual and supervised labeling. This underscores the result of increasing medical image analysis. Overall, the PadChest dataset is widely used as a crucial tool for researchers in medical imaging.

**A New Web-Supervised Method for Image**

**Dataset Constructions**

This literature conducted a study about image dataset collection using the web-supervised method. They aim to create a model that automatically collects images from the internet based on the given queries. The researchers proposed methodology and system framework are query expanding, filtering noisy query expansions, visual-non salient expansion filtering, artificial images filtering, clustering-based noisy image filtering, and CNN-based noisy image filtering [11]. In their experiment, the researchers concluded that the image dataset collected by the model has shown better results. The evaluation indicated that the created AutoImgSet-10 dataset presented a higher average accuracy in object detection tasks. This finding also validates the efficiency of the proposed structure for creating image datasets.

**DatasetGAN: Efficient Labeled Data Factory with Minimal Human Effort**

This study focuses on their DatasetGAN model, which can generate many datasets requiring minimal human intervention. The researchers' approach in making the model work is to obtain a few images of the object, which will be mapped for its features. After that, a person will be able to label the images with the chosen set of labels. Lastly, the model is trained to match the labels provided [12]. After experimenting, the researchers' model proved useful and powerful, with only a handful of annotated images, which allowed the model to learn from this method. This study examined the model's ability to effectively use the acquired hidden structures from GANs to create high-quality images with semantic categorization. This was achieved using only a small number of manually labeled examples. Therefore, the advancement offers opportunities for generating large-scale datasets in computer vision applications.

III. METHODOLOGY

1. Present and discuss the dataset used in this experiment and its characteristics.

The Garbage Image Dataset was created to understand waste categorization using deep learning models. It includes four categories (Cardboard, Metal, Paper, and Plastic). Moreover, the dataset includes 350 images in each category and a total number of 1,400 images. The objective is to help individuals correctly classify waste and enable industry to achieve environmental sustainability. It is also a valuable tool for conducting experiments and evaluations in image classification, which primarily focuses on garbage-sorting methods. Therefore, it is essential for studies that enhance processes for accurately classifying waste items to improve waste management strategies.

1. Explain the preprocessing steps applied to clean and prepare the dataset for MlogReg modeling

Initially, the selected photos were converted into a standardized format, such as JPEG, to ensure consistency in their display. Each image was labeled accurately and assigned to its corresponding category in alignment with the total number of images. Afterwards, the annotations for the images were merged into a unified folder, making it easier to access them from Colab to the Google Drive repository. The following steps involved resizing the images and transforming them from color to grayscale, improving their suitability for uniform input into the model. It is necessary to ensure that the dataset implemented is a balanced dataset pertaining to it should have the equal amount of collected images per category, which is a crucial factor for fostering a comprehensive and unbiased prediction model. In regards with the Google Drive repository, two folders were organized: one for the collected images and an empty folder reserved for the automatically stored resized data upon code initialization. This preprocessing pipeline ensures that the dataset is appropriately formatted, labeled, balanced, and is ready for effective utilization in Multinomial Logistic Regression modeling.

1. Discuss the details of the initial CNN architecture that will be used for the image classifier model as well as the performance metrics that will be applied.

For the initial configuration of our CNN architecture, we've opted for simplicity to establish a baseline. Our model consists of two convolutional layers, essential for learning hierarchical features, a max pooling layer, and an output layer that represents the labels from the garbage image dataset, which utilizes a softmax activation function for multi-class classification, etc. A hidden layer containing 128, 64, 190 neurons with a rectified unit linear unit as its activation function will also be included to capture the patterns and train the model. To prevent overfitting, we've also included a dropout layer with a 0.2 dropout rate as its starting value. The initial architecture will be modified one by one as new models are developed. This approach will allow us to identify which configurations best fit the model. We've chosen accuracy as our primary performance metric for the given size of our dataset, having exactly 350 images per category. As we progress in the experiment, we will systematically explore various refinements to find the optimal configuration that aligns with the characteristics of our dataset. This iterative process will guide us toward a tailored CNN architecture that maximizes accuracy and finds the top-performing model.

1. Discuss the strategy formulated by the group to adjust the different hyperparameters of the CNN to get the best model. Note: There is no limit to the hyperparameters that will be covered in this exercise.

Regarding our team's strategy, our approach to achieving optimal performance in both the holdout and k-fold validation involves utilizing isolation. The utilization of this strategy has demonstrated its efficiency for us, as it enables an in-depth examination of the particular effects of hyperparameters on accuracy, ensuring that none of the factors that influence performance are overlooked. The thorough examination allows our research group to make well-informed decisions and employ insights derived from the outcomes of the arbitrary model as an additional approach.

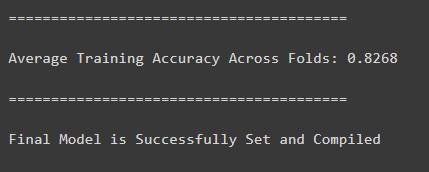
This applies approximating initial values for hyperparameters by assessing the results of initial model iterations. Our team analyzes various values for each hyperparameter through multiple experiments, carefully evaluating their effects on the model's overall accuracy. This technique allows us to achieve an optimal balance between underfitting and overfitting. We maintain a complete record in an Excel file to improve our strategy. This record is an informative resource, allowing an organized approach that enables our team to identify which aspects of the hyperparameters require modification quickly. By implementing this strategy, we have successfully achieved a good performance model and proven the effectiveness of our approach in this study.

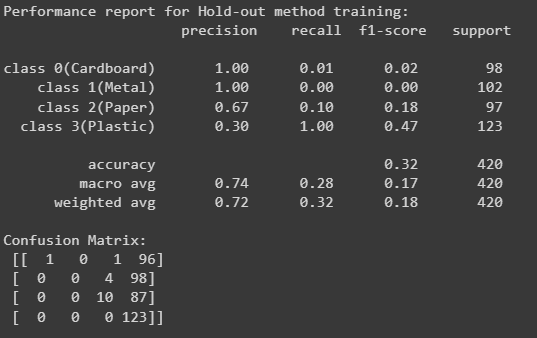
1. Discuss the format of the Excel file that you will use to record the results of each run in your experiment.

This Excel file documented our experiment results for multiple models and various hyperparameters. The parameters include important factors such as the number of filters that determine the value of collected features, the pooling size which determines the level of summarization, kernel size that defines the parameters of the extracted features, dropout rates that control the level of normalization, epochs that represent the number of repetitions through the dataset throughout training, batch sizes that refer to the number of samples used for each training, learning rates that adjust the scale in the optimization technique, neurons in hidden layers that represent the nodes in intermediary layers of the network, data augmentation that develops the training dataset through statistical changes, k-folds training accuracy that assesses outcomes through cross-validation, and holdout method that defines the performance accuracy measures the model's performance on unknown data. The collecting data from different Excel files can provide a detailed representation of the tests that were conducted. As a result, this collecting data greatly improved the group's understanding of the model outcomes. This evaluation also allows an assessment of the impact of different hyperparameters that help determine the model that achieves the highest efficiency and performance.

IV. RESULTS

1. Present and discuss the results of your experiments.

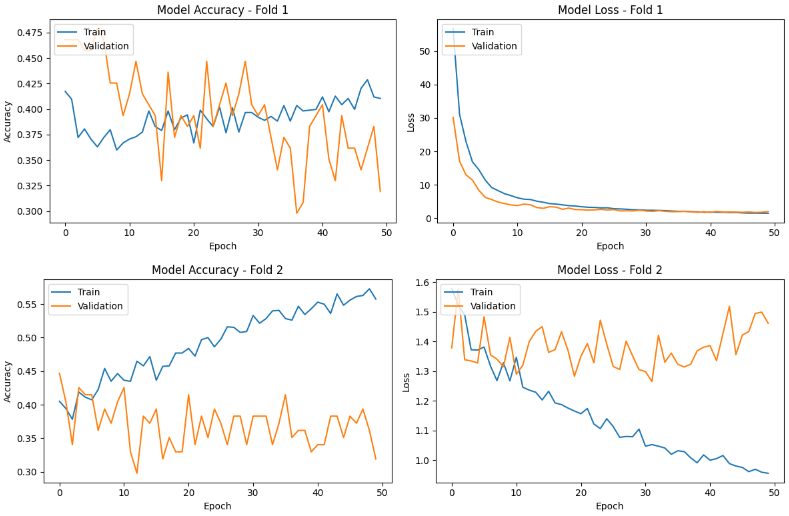


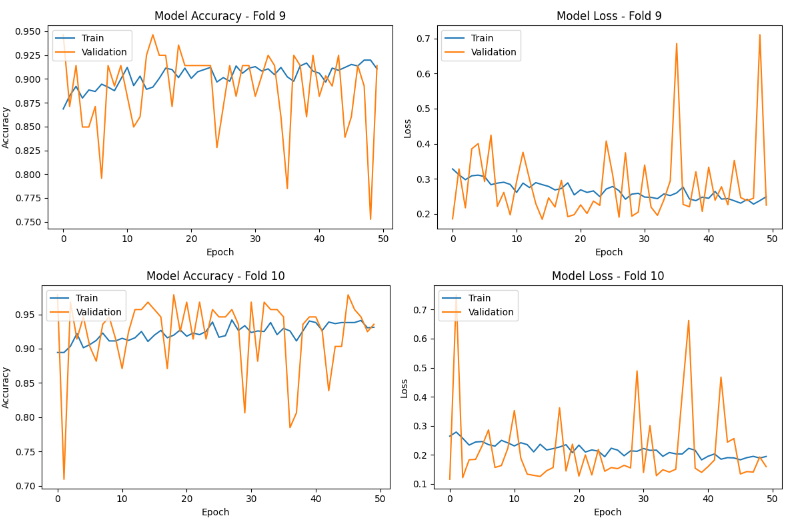
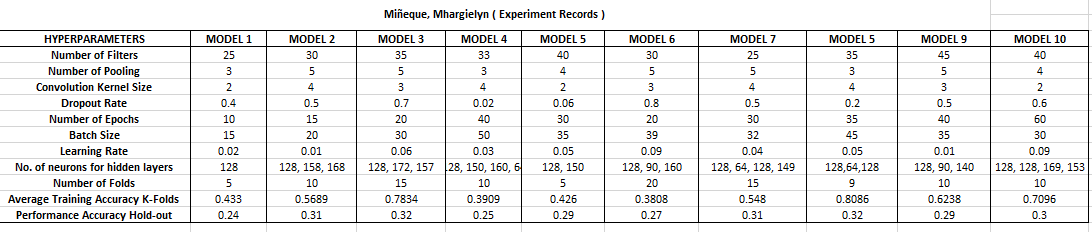


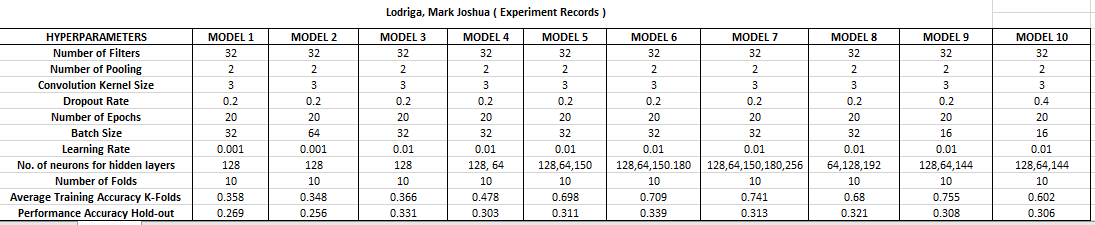
*Figure 1: Accuracy Result of K-fold and Hold-out Method*

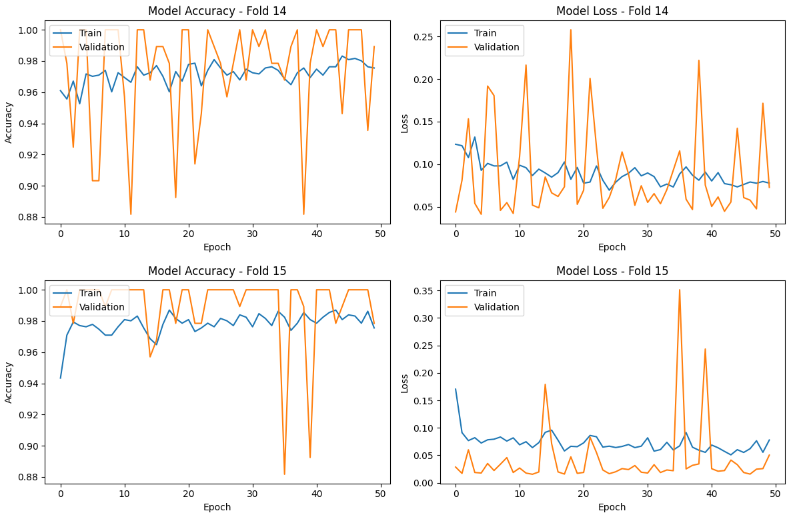
Based on the overall result that we obtained, the average result accuracy of the K-fold was 0.8268 or 82.68%, and the hold-out resulted in only 0.32 or 32%. For the cross-validation using Stratified K-Fold, it initializes 15 folds along with the configuration of 50 filters, a pooling size of 4, a convolution kernel size of 5, and a dropout rate of 0.3. The training process initializes 50 epochs, the same as the batch size of 50, and a learning rate is set at 0.08. The hidden layers of the neural network consist of three, along with the values 128, 64, and 190 for neurons. In this method, the model reached an accuracy of 82.68%, which demonstrates its ability to generalize well-unrecognized data. On the other hand, the hold-out validation approach with the same configuration of hyperparameters exhibited an average accuracy of 32%. This metric provides valuable insights into the model's result across different subsets of the dataset. The lower average accuracy in this method compared to the k-fold suggests that the model might encounter challenges in generalizing to diverse data distributions. Further analysis and refinement of the hyperparameters could be examined to enhance the model's overall performance and reliability across various data partitions. Overall, this experiment indicates that the kfold validation provides a more consistent and higher performance during the training process. This method signifies the suitability and adaptability to various data distributions, allowing the model to capture patterns associated with different types of waste and leading to a more reliable classification method.

1. Display and analyze the performance charts of the top-performing model in both the hold-out and K-fold training process







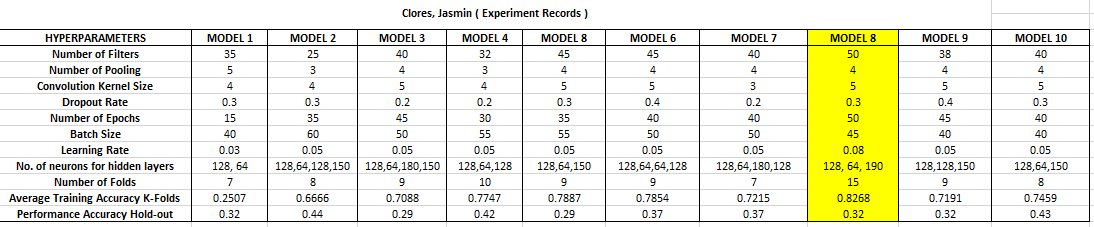


*Figure 2: Performance Chart for Training and Validation Accuracy & Loss*

As seen in the figure, the performance of both the training and validation datasets is consistent for most of the folds across all 50 epochs. This result suggests that the model learned and can generalize images, indicating that it is able to identify objects even when presented with new data. The diagrams for how much the model lose during training also back this up for most of the graph; the gaps between the losses for the training and validation sets are small. This is a good sign as it indicates that the model can correctly identify unseen data frequently. Although the accuracy performance for most of the folds looks good, there are some unusual instances of accuracy, as seen in folds 14 and 15. There are instances where the validation accuracy reaches perfect accuracy multiple times across all the epochs; this occurrence may suggest overfitting for these specific folds.

V. ANALYSIS

1. Interpret the results based on your observation.



*Figure 3: Experiment Records Result*

Figure 3 represents the performance of 10 different models made by each team member. Model 8 of Jasmin Clores stands out with a high accuracy value. By adjusting different hyperparameters, we achieved an average accuracy of 0.8268 using K-fold validation. This result indicates that the model has effectively learned from the training data and can make accurate predictions that demonstrate a significant level of generalization. Another method is hold-out validation, which indicated a lower mean accuracy of 32%. The lowered accuracy in hold-out validation suggests a potential limitation in the model's performance when presented with new, unknown data. Analyzing these findings is crucial in order to determine the most suitable approach for our project proposal. The difference in K-fold and hold-out validation accuracies involves a more thorough analysis of the utilized data and models. Therefore, this experiment result will more ensure that the selected methodology accurately assesses the model's actual performance in real-life situations.

1. Discuss any challenges or limitations you encountered during the experiments.

One of the challenges that our group encountered was adjusting the different hyperparameters. In this case, the model has two different methods, which means there are a lot of hyperparameters that we need to monitor either increasing or decreasing to develop the top-performing model. When we increase the epochs and K-fold method, it takes a lot of time to wait for the result of the model. It was also a challenge for us to achieve acceptable accuracy in training and testing for K-fold and hold-out, which we can monitor if the mode is underfitting or overfitting. This experiment encourages us to conduct a broader study of combined approaches with reliable cross-validation methods that develop an increased awareness of data points and model challenges. Therefore, these challenges enable us to develop a good prediction model regarding our chosen domain.

VI. CONCLUSIONS

In conclusion, this paper on Image Classification utilizing Custom Images with K-fold validation is a useful tool for better understanding our project proposal. Our group also explored different experiments and learned insights to adjust hyperparameters to improve the efficiency of the proposed model and achieve the best model performance result. This activity will help us determine whether the hold-out or K-fold method would be most suitable for our project proposal. It also provides information about the practical use of image classification algorithms to allow us to gain a greater understanding of how these methods align with organized proposed project objectives. Furthermore, this study significantly contributes to our group's ability to prepare a detailed project proposal for the course to provide foundational knowledge and methodologies for the project development.

# REFERENCES

1. “Creating a custom dataset,” Slate Knowledge Base, Aug. 05, 2023. <https://knowledge.technolutions.com/hc/en-us/articles/16128518234267-Creating-a-Custom-Dataset>
2. S. Sangidwar, “Dataset Generator: Creating And Analyzing Customized Data,” 2023. <https://www.semanticscholar.org/paper/Dataset-Generator%3A-Creating-And-Analyzing-Data-Sangidwar-Mane/413b9b052687be30809efe399cdcb1ea24632b6a>
3. G. Horsman and J. R. Lyle, “Dataset construction challenges for digital forensics,” Forensic Science International: Digital Investigation, vol. 38, p. 301264, Sep. 2021, doi: 10.1016/j.fsidi.2021.301264.
4. “What is Custom Dataset,” Tasq.ai, Nov. 23, 2022. https://www.tasq.ai/glossary/custom-dataset/
5. J. Allibhai, “Hold-out vs. Cross-validation in Machine Learning - Jaz Allibhai - Medium,” Medium, Jun. 21, 2022. [Online]. Available: https://medium.com/@jaz1/holdout-vs-cross-validation-in-machine-learning-7637112d3f8f
6. “Analysis of k-Fold Cross-Validation over Hold-Out Validation on Colossal Datasets for Quality Classification,” IEEE Conference Publication | IEEE Xplore, Feb. 01, 2016. https://ieeexplore.ieee.org/document/7544814
7. “Custom Dataset Creation with Tensorflow Framework and Image Processing for Google T-Rex,” IEEE Conference Publication | IEEE Xplore, Sep. 25, 2020. https://ieeexplore.ieee.org/document/9242565
8. N. C. Mithun, R. Panda, and A. K. Roy-Chowdhury, “Generating Diverse Image Datasets with Limited Labeling,” ResearchGate, Oct. 2016, doi: 10.1145/2964284.2967285.
9. R. Duan, H. Deng, M. Tian, Y. Deng, and J.-R. Lin, “SODA: A large-scale open site object detection dataset for deep learning in construction,” Automation in Construction, vol. 142, p. 104499, Oct. 2022, doi: 10.1016/j.autcon.2022.104499.
10. A. Bustos, A. Pertusa, J. M. Salinas, and M. De La Iglesia-Vayá, “PadChest: A large chest x-ray image dataset with multi-label annotated reports,” Medical Image Analysis, vol. 66, p. 101797, Dec. 2020, doi: 10.1016/j.media.2020.101797.
11. Y. Yao, J. Zhang, F. Shen, X.-S. Hua, J. Xu, and Z. Tang, “A new web-supervised method for image dataset constructions,” Neurocomputing, vol. 236, pp. 23–31, May 2017, doi: 10.1016/j.neucom.2016.07.066.
12. Y. Zhang, “DatasetGAN: Efficient Labeled Data Factory with Minimal Human Effort,” arXiv.org, Apr. 13, 2021. https://arxiv.org/abs/2104.06490

‌