Received: 05 March 2022 | *Revised: 27 April 2022* | *Accepted: 27 April 2022*
DOI: 10.37917/ijeee.18.2.2 *Vol. 18| Issue 2| December 2022*

Open Access

Iraqi Journal for Electrical and Electronic Engineering
Original Article

Shapley Value is an Equitable Metric for Data Valuation

Seyedamir Shobeiri*, Mojtaba Aajami
Department of Computer Engineering, Islamic Azad University of Zanjan, Zanjan, Iran

Correspondence
*Seyedamir Shobeiri
Department of Computer Engineering,
Islamic Azad University of Zanjan, Zanjan, Iran
Email: seyedamir.shobeiri@iauz.ac.ir

Abstract
Low-quality data can be dangerous for the machine learning models, especially in crucial situations. Some large-scale datasets have low-quality data and false labels, also, datasets with images type probably have artifacts and biases from measurement errors. So, automatic algorithms that are able to recognize low-quality data are needed. In this paper, Shapley Value is used, a metric for evaluation of data, to quantify the value of training data to the performance of a classification algorithm in a large ImageNet dataset. We specify the success of data Shapley in recognizing low-quality against precious data for classification. We figure out that model performance is increased when low Shapley values are removed, whilst classification model performance is declined when high Shapley values are removed. Moreover, there were more true labels in high-Shapley value data and more mislabeled samples in low-Shapley value. Results represent that mislabeled or poor-quality images are in low Shapley value and valuable data for classification are in high Shapley value.

KEYWORDS: Artificial intelligence, Machine learning, Shapley Value, Black box, Datasets.

I. INTRODUCTION

Machine learning models methods such as deep learning are in dire need of data on a large scale to have high accuracy and they have incredible performance in the medical area, including skin lesion classification from dermatoscopy [1], automated chest X-ray interpretations [2] and intracranial hemorrhage detection from computed tomography [3]. large-scale training datasets hand-labeled is available to achieve this success. but, hand labeling of large-scale datasets is boring, time-intensive and in the medical part it is needed expertise [4]. on the other hand, crowd-sourcing or automated algorithms is emerged to label huge datasets. for example, MIMIC-CXR [5], chest X-ray [6] and DeepLesion [7]. so, the results indicate these methods are more accurate than hand-labeled datasets [8]. Moreover, probably datasets contain inaccurate labels in addition datasets could contain noise and different types of artifacts error [9]. Likewise, the value of data is an important challenge in datasets, especially in large-scaler [10]. machine-learning models' accuracy is decreased when they are trained on datasets containing low-quality data [11]. So, algorithms that are able to automatically recognize low-quality data could improve this problem. In this work, we suggest using Shapley Value [12] to identify low-quality data in ImageNet datasets [13]. there are some ways to manage inaccurate labels such as data re-weighting [14] and adding a noise layer in the network architecture [15]. for instance, label noise is managed for the ChestX-ray dataset [6]. previous studies focus on handling suboptimal images or inaccurate labels in training stages or model development. but Shapley Value directly recognizes low-quality data, improves ML model. Given a supervised learning algorithm, a training set, and a predictor performance score, Shapley Value [12] is a metric that calculates the value of each training data to the predictor performance. examine on small to moderate-scale, imaging and synthetic data have indicated that low Shapley value captures, while high Shapley value represents the type of new data that should be acquired to most efficiently improve the predictor performance [12], In addition, Shapley Value has better performance than the leave-one-out (LOO) score [16].

furthermore, Shapley value indicates several benefits as a framework for data valuation [12]: 1. Natural properties of equitable data valuation are satisfied by the Shapley Value. 2. Shapley Value assigns a single value score for each data point so directly interpretable are enabled. our goal is to find the effectiveness of data Shapley in capturing low-quality data as well as informing valuable data in the context of classification from ImageNet images.

Our primary contributions are as follows:

a) first of all, A framework (see Fig. 1) is expanded to evaluate the value of training data in a large ImageNet

dataset in the context of classification by using data Shapley values. b) we represent that high Shapley Value indicates data points that are valuable examples, while low Shapley Value displays mislabeled for classification.

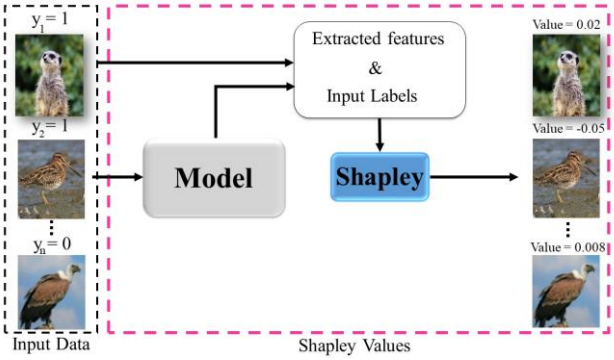

Fig. 1: Overview of Shapley Value method

Figure 1 is explained: First, the input data were ImageNet images and their corresponding labels from the ImageNet dataset. Second, to compute data Shapley values for the training data, we first extracted feature vectors from a pre-trained convolutional neural network (CNN), ImageNet. then, we applied Sample Value to approximate the value of each training point, and the predictor performance score was prediction accuracy for classification.

II. Material and Methods

A. *Shapley Value*

Shap Value performs the equitable data valuation in supervised machine learning. For a given set of training data points D and a performance metric, the "Shapley Value" value of a data point xi ∈ D is defined as:

$$\emptyset = \sum_{SCD-\{i\}} \frac{V(s\cup\{i\})-V(S)}{\binom{n-1}{|S|}} \quad (1)$$

Where V(S) is the performance of the model trained on subset S of the data. V(S)is the prediction accuracy on the validation set. Intuitively, the Shapley value of a data point is a weighted average of its marginal contribution to subsets of the rest of the dataset. As a result, it can be used as a measure of data quality: a data point with a high Shapley value is one that improves the model's performance if we add it to most subsets of the data, while a data point with a negative value on average hurts the performance of the model. Exact computation of Eq. requires an exponential number of computations in the size of the dataset, which is infeasible in most realistic settings. In fact, High value indicate high quality of image and correct label while low value represents low quality of image and incorrect label. Finally, SHAP Value has three outputs, which are value, growth rate, and main data, respectively. Value: An array that each cell represents a pixel, each cell of the array contains another array that contains three cells and represents the RGB effect as shown: Value = [[R, G, B], [R, G, B], [R, G, B], …] and the growth rate, which is a base number, and our main data, which is the original values of our image. How to calculate the value of an image is as follows:

$$Value = \sum_{i=0}^{n}([R_i + G_i + B_i])) + Base \quad (2)$$

According to the above formula i = 0 because the array cells start from zero and N is the number of pixels, R represents the effect of red, G represents the effect of green, B represents the effect of blue, which indicates each of these pixel colors. How effective the image has been in our model is that the sum of the effects of red, green, and blue with the base, which represents the growth rate, reflects the value of image. next, two datasets are tested that whether the value of an image is always the same or depends on another factor. Shapley Value library is called SHAP in python. In this paper, we used it to interpret the value of the ImageNet dataset. The blue pixel indicates the low quality of the image that this pixel affected on machine learning algorithms to choose other classes, on the opposite, the red pixel represents the high quality of data that affected choosing a true class.

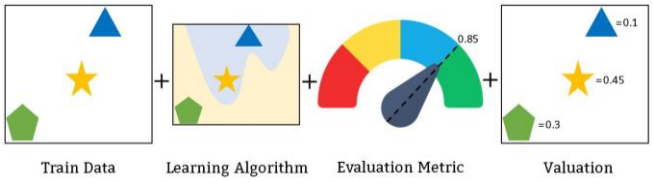

Fig. 2: Shapley Value

As you can observe in Fig. 3 there are real images on the left side. There are the outputs of Shapley value on the right and above them, there are the labels of each output. A degree is provided at the bottom. value of red color represents positive efficiency to choose this image for this label and on the other hand, the blue color indicates negative efficacy to choose it in other class.

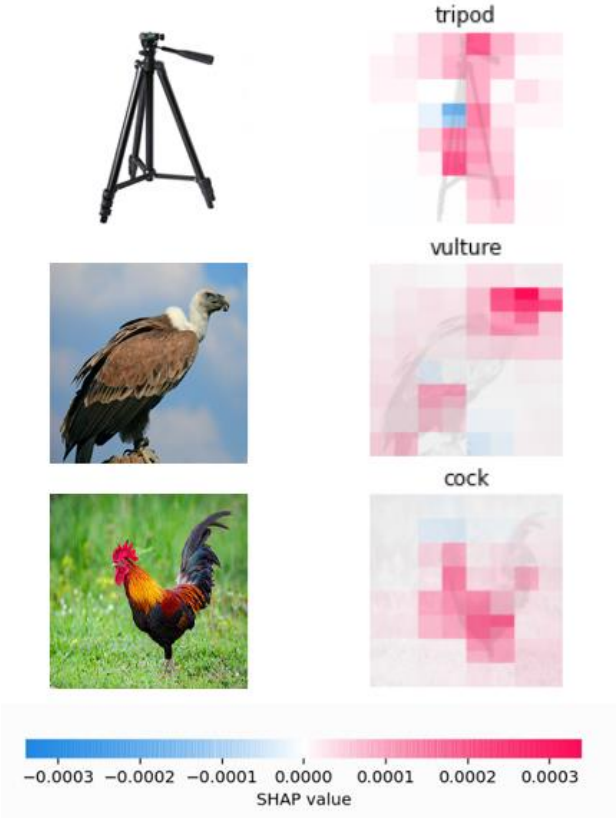

Fig. 3: Shapley Value output Samples for images.

B. *Leave_one_out (LOO)*

Leave_one_out or LOO is a method for validation data.in this method, the dataset is split to train and test data in the amount of k-fold-1, and this method is repeated in the amount of k [16].

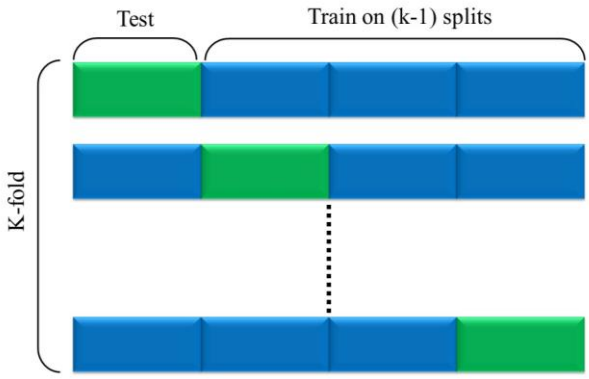

Fig. 4: Leave_One_Out (LOO)

C. *Random method*

Python programming language consists of huge libraries that help researchers and programmers reach their goals easily. python has a built-in library for random numbers. This library has a lot of functions. Randint function is used in the paper. Randint function return discrete between the range of numbers you choose.

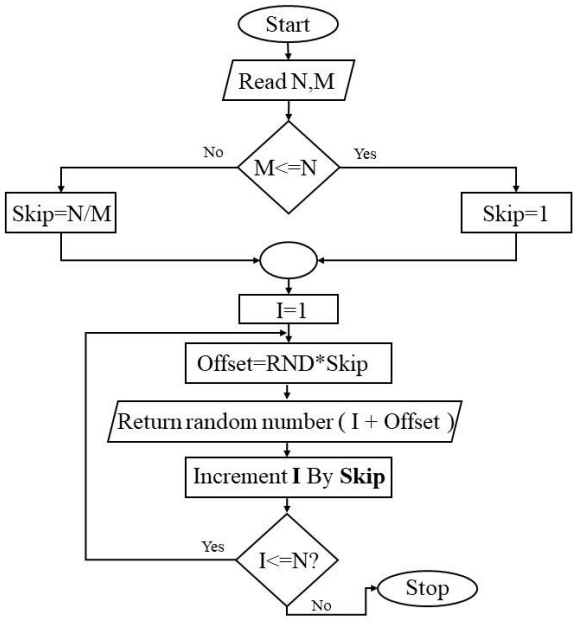

Fig. 5: Random method

D. *ImageNet Datasets*

ImageNet dataset is used in this paper. this dataset is a large scale which means it contains 14,197,122 images with 1000 classes. This dataset is the fundamental base of objects recognition [13]. Fig. 6 shows an example of objects classification or ImageNet dataset [13].

Fig. 6: Sample of ImageNet dataset

E. *Data Analyzing*

In this paper, Accuracy, and recall of the training and testing model is used with python language. A confusion matrix is a summary of classification prediction results.it shows correct and incorrect predictions and is broken down by each class [17].

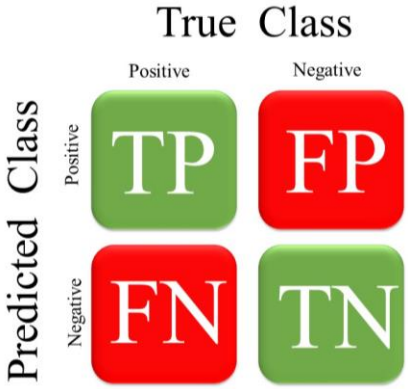

Fig. 10: Confusion matrix

we calculate accuracy metrics on Shapley value, LOO and random when we remove high and low-quality data. The accuracy mathematical formula is shown in equation (3).

$$\text{Accuracy} = \frac{(TP+TN)}{(TP+FP+TN+FN)} \qquad (3)$$

TP implies the amount of positive data that the model predicts as positive, TN implies the amount of negative data that the model predicts as negative, FP implies the amount of negative data that the model predicts as positive, and FN implies the amount of positive data that the model predicts as negative [18].

II. The Results

In this paper, our goal is to explain the effectiveness of data Shapley in recognizing valuable and low-quality data in a large public ImageNet dataset. features are extracted from a pre-trained convolutional neural network (CNN) called VGG16, and computed the data Shapley value of each training point with respect to the accuracy of a logistic regression algorithm for classification. in addition, in collaboration with another colleague, the least valuable and most valuable are evaluated for classification in the

ImageNet dataset. Qualitative interpretations for their Shapley values are provided in the next section. Important data points are recognized by Shapley Value for classification. As figures represent training data. training images contain 42.5% of the negative Shapley Values. after evaluating the Shapley Value, data datums are removed, and each time when 1% of the training data were removed a new logistic regression model is trained. Data points were removed and figures indicate the changes in prediction accuracy. model performance is declined when high Shapley Value data points are removed. In contrast, removing randomly data points or removing high LOO values data points had a small effect on the performance of the model.

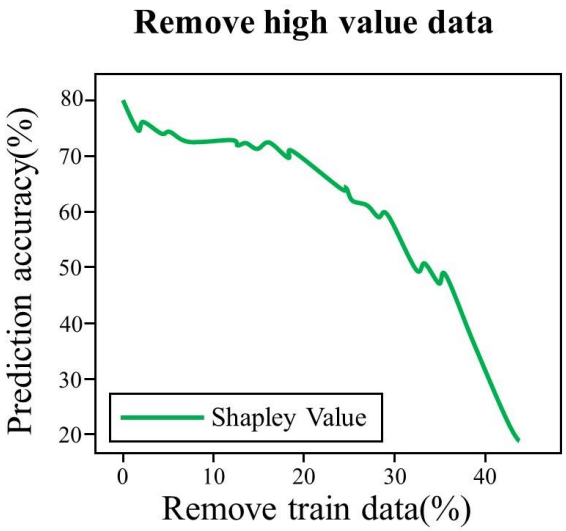

Fig. 11: Result of testing to find accuracy

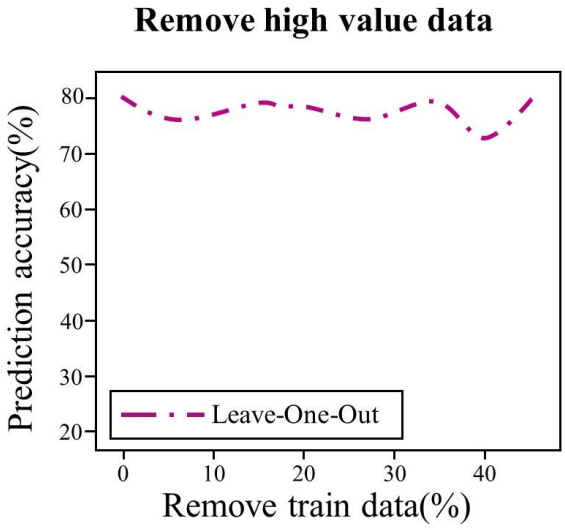

Fig. 12: Result of testing to find accuracy

these are represented in the data Shapley Value is a highly accurate measure of a data point's importance since data points with high Shapley values were crucial to classification. The Mislabels in the dataset are identified by low Shapley Value. My senior data science colleague re-label the 100 high-quality and 100 low-quality data and 100 randomly sampled ImageNet images in the training set. There were important observations. First, there were many more mislabels in low value images. There were important observations. First, many more mislabels existed in low-value images. observations are indicated many more mislabels in low-quality images are existed to compare randomly sampled or high value. there is a relation between the Shapley values and image quality. Data scientist reported that all of the 200 ImageNet images met diagnosis quality. but there were nine images where a portion of the object field was out of the image frame. Among these nine images, seven had negative Shapley values whether or not they were correctly or incorrectly labeled. Whereas the other two images had positive Shapley values and were correctly labeled. Therefore, this suggests that low Shapley values not only indicate mislabels but also poor image quality.

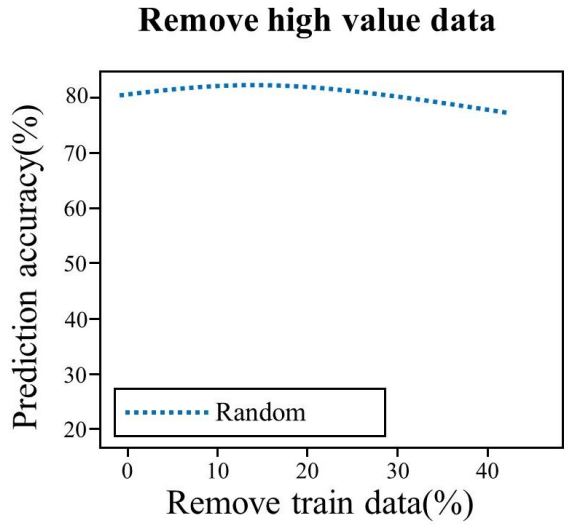

Fig. 13: Result of testing to find accuracy

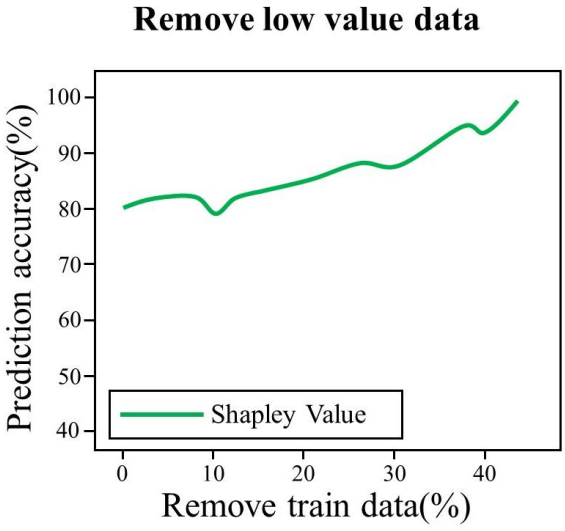

Fig. 14: Result of testing on to find accuracy

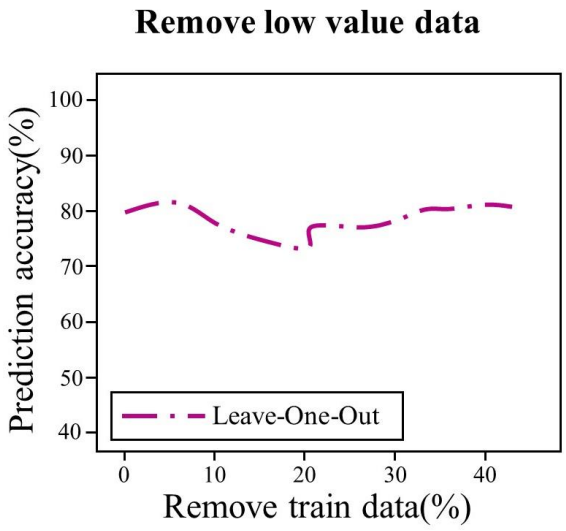

Fig. 15: Result of testing on to find accuracy

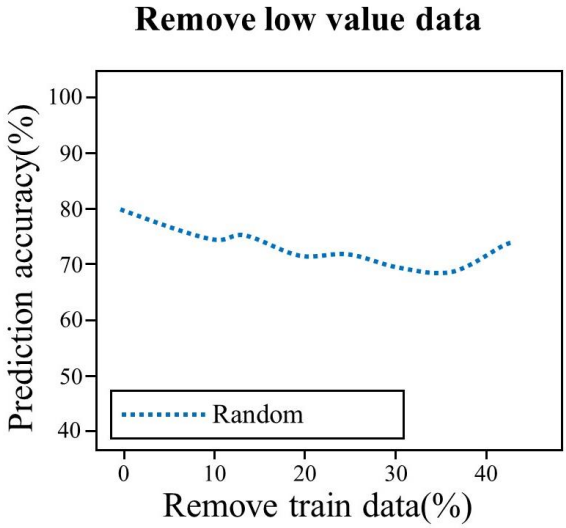

Fig. 16: Result of testing on to find accuracy

IV. Conclusion

Machine-learning algorithms are practical in real life, especially in medical and healthcare. therefore, ML algorithms should be reliable that fortunately, Shapley Value helps ML to be reliable in crucial situations. This prediction method will be accurate and it can show what happens in ML black box and convert it to a white box and interpret the ML algorithm. It can show mislabeled data, high and low-quality data when you want to pay the price of data for data individuals' generator. In other words, it can be called the equitable valuation data method.

Conflict of Interest

The authors have no conflict of relevant interest to this article

References

[1] A. Rezvantalab, H. Safigholi, and S. Karimijeshni, "Dermatologist Level Dermoscopy Skin Cancer Classification Using Different Deep Learning Convolutional Neural Networks Algorithms," arXiv:1810.10348.https://doi.org/10.48550/arXiv.1810.10348

[2] P. Rajpurkar *et al.*, "Deep learning for chest radiograph diagnosis: A retrospective comparison of the CheXNeXt algorithm to practicing radiologists," *PLoS Med*, vol. 15, no. 11, p. e1002686, Nov. 2018.

[3] J. J. Titano *et al.*, "Automated deep-neural-network surveillance of cranial images for acute neurologic events," *Nat Med*, vol. 24, no. 9, pp. 1337–1341, Sep. 2018.

[4] N. Noori and A. Yassin, "Towards for Designing Intelligent Health Care System Based on Machine Learning," *IJEEE*, vol. 17, no. 2, pp. 120–128, Dec. 2021.

[5] A. E. W. Johnson *et al.*, "OPEN MIMIC-CXR, a de-identified Data Descriptor publicly available database of chest radiographs with free-text reports," *Scientific Data*, p. 9.

[6] X. Wang, Y. Peng, L. Lu, Z. Lu, M. Bagheri, and R. M. Summers, "ChestX-ray8: Hospital-Scale Chest X-Ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases," p. 10.

[7] K. Yan, X. Wang, L. Lu, and R. M. Summers, "DeepLesion: Automated Deep Mining, Categorization and Detection of Significant Radiology Image Findings using Large-Scale Clinical Lesion Annotations," *arXiv:1710.01766 [cs]*, Oct. 2017, Accessed: Apr. 23, 2022.

[8] L. Oakden-Rayner, "Exploring large scale public medical image datasets," *arXiv:1907.12720 [cs, eess]*, Jul. 2019, Accessed: Apr. 23, 2022.

[9] M. J. Willemink *et al.*, "Preparing Medical Imaging Data for Machine Learning," *Radiology*, vol. 295, no. 1, pp. 4–15, Apr. 2020.

[10] O. Diaz *et al.*, "Data preparation for artificial intelligence in medical imaging: A comprehensive guide to open-access platforms and tools," *Physica Medica*, vol. 83, pp. 25–37, Mar. 2021.

[11] J. R. Zech, M. A. Badgeley, M. Liu, A. B. Costa, J. J. Titano, and E. K. Oermann, "Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study," *PLoS Med*, vol. 15, no. 11, p. e1002683, Nov. 2018.

[12] S. Shobeiri and M. Aajami, "Shapley value in convolutional neural networks (CNNs): A Comparative Study," *American Journal of Science & Engineering*, vol. 2, no. 3, pp. 9–14, Dec. 2021.

[13] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A Large-Scale Hierarchical Image Database," p. 8.

[14] C. Xue, Q. Dou, X. Shi, H. Chen, and P. A. Heng, "Robust Learning at Noisy Labeled Medical Images: Applied to Skin Lesion Classification," *arXiv:1901.07759 [cs]*, Jan. 2019, Accessed: Apr. 24, 2022.

[15] Y. Dgani, H. Greenspan, and J. Goldberger, "Training a neural network based on unreliable human annotation of medical images," in *2018 IEEE 15th International*

Symposium on Biomedical Imaging (ISBI 2018), Washington, DC, Apr. 2018, pp. 39–42.
[16] R. D. Cook, "DETECTION OF INFLUENTIAL OBSERVATIONS IN LINEAR REGRESSION," p. 13.
[17] V. M. Patro and M. Ranjan Patra, "Augmenting Weighted Average with Confusion Matrix to Enhance Classification Accuracy," *TMLAI*, vol. 2, no. 4, Aug. 2014.
[18] S. García, A. Fernández, J. Luengo, and F. Herrera, "A study of statistical techniques and performance measures for genetics-based machine learning: accuracy and interpretability," *Soft Comput*, vol. 13, no. 10, pp. 959–977, Aug. 2009.