DATA SHAPLEY VALUE FOR HANDLING NOISY LABELS: AN APPLICATION IN SCREENING COVID-19 PNEUMONIA FROM CHEST CT SCANS

Nastaran Enshaei[†], *Moezedin Javad Rafiee, MD*[‡], *Arash Mohammadi*[†], *and Farnoosh Naderkhani*[†]

[†]Concordia Institute for Information Systems Engineering, Concordia University, Montreal, Canada
[‡]Department of Medicine and Diagnostic Radiology, McGill University, Montreal, QC, Canada
[5]Department of Electrical and Computer Engineering, University of Toronto, Toronto, Canada

ABSTRACT

A long-standing challenge of deep learning models involves how to handle noisy labels, especially in applications where human lives are at stake. Adoption of the data Shapley Value (SV), a cooperative game-theoretic approach, is an intelligent valuation solution to tackle the issue of noisy labels. Data SV can be used together with a learning model and an evaluation metric to validate each training point's contribution to the model's performance. The SV of a data point, however, is not unique and depends on the learning model, the evaluation metric, and other data points collaborating in the training game. However, effects of utilizing different evaluation metrics for computation of the SV, detecting the noisy labels, and measuring the data points' importance has not yet been thoroughly investigated. In this context, we performed a series of comparative analyses to assess SV's capabilities to detect noisy input labels when measured by different evaluation metrics. Our experiments on COVID-19-infected of CT images illustrate that although the data SV can effectively identify noisy labels, adoption of different evaluation metric can significantly influence its ability to identify noisy labels from different data classes. Specifically, we demonstrate that the SV greatly depends on the associated evaluation metric.

***Index Terms*: Data Shapley value, Noisy Labels, Data Valuation, Medical Imaging, Capsule Networks.**

1. INTRODUCTION

Deep learning has proven remarkable success in several medical fields, including medical imaging, where advanced computer vision algorithms have delivered near human-level performance in specific tasks. Several factors may affect the success of a deep learning algorithm, including model structure, initialization of the model parameters, training methods, and computational hardware. However, the most influential factor is having access to large-scale data sets with reliable labels. Collecting large-scale data sets with reliable labels is, however, a significantly challenging task in several applications, particularly within the medical domain. On the one hand, patients' privacy preservation and data sharing protocols prohibit hospitals and clinical institutions from releasing their in-house data sets. On the other hand, labeling medical images requires knowledge and expertise from radiologists and physicians, making the labeling process resource-intensive. To overcome this challenge, extracting pathology labels of large-scale medical images from radiology reports with the help of text mining [1], human-machine collaborative techniques [2, 3], and the use of non-expert annotators are some of the proposed solutions in the literature [4]. However, the provided labels by such methods are noisy and inaccurate compared to manually labeled images [5]. Consequently in this context, handling noisy labels is crucial for developing high-performance models.

Prior work: Learning from a training data set with noisy labels has been a challenging task within the deep learning domain for a long time. Research studies indicate that noisy labels may have a more significant adverse effect on the performance of deep learning models than the noises in data attributes/measurements [6]. Proposed techniques for handling noisy labels in medical imaging data sets include weakly supervised learning [7–9], customized training methods [4], and re-weighting training samples [10]. Another approach to tackle the noisy label problem is to adopt Shapley Value (SV), a cooperative game-theoretic method, where each training point is considered a player in the training game, and its contribution to any subset of players is measured using a performance evaluation metric. Researchers have leveraged the SV for data valuation and measuring the quality of data points in different deep learning applications [11–18]. In [19] and [20], authors utilized the SV in a federated learning study to measure the quality of data from each participant. A recent study [17] has investigated the capability of the SV in quantifying the importance of training data points in the performance of the learning model in a large-scale X-ray data set. It should be noted that the SV of a data point is not a unique value and depends on the learning model, the evaluation metric, and other data points collaborating in the training game. So far, research studies have used only one specific evaluation metric in their SV computation process, such as accuracy metric in [17, 19] and Area under the ROC curve (AUC-ROC) metric in [20] to demonstrate the data SV capability in measuring the quality of data points. However, the effect of utilizing different evaluation metrics in computing the SV, detecting the noisy labels, and measuring the data points' importance in the training process has never been investigated.

Contributions: This study investigates effects of incorporating different evaluation metrics in determining the data SV and quantifying the importance of each training point in the model's performance. To the best of our knowledge, this is the first study that explores adoption of various evaluation metrics in measuring the data SV and detecting noisy input labels. The data SV measures the contribution of each training point to all possible subsets of the training set, having an exponential complexity in the size of the training set. Therefore, we use a permutation sampling approach to approximate the data SV. We assess the capability of data SV obtained by commonly-used classification metrics, including accuracy, recall, and specificity in detecting the noisy labels through a set of experiments with different noise levels. The experiments are performed based on a COVID- 19 screening task from chest CT scans by implementing a lightweight Capsule-network-based classifier [21] to extract discriminative features from chest CT scans and distinguish COVID-19-infected CT images from normal ones [22]. The Capsule-network-based classifier can represent each CT image via a small feature vector. Since calculating the SV requires retraining the learning model on multiples coalitions of training samples, we use a fast classifier to make

978-1-6654-0540-9/22/$31.00 ©2022 IEEE

our data valuation process time-efficient. Therefore, the extracted feature vector is fed into a logistic regression classifier for final decision making. The results indicate that the measured data SV is not unique and is highly dependent on the evaluation metric. We demonstrate that despite the great potential of the data SV in detecting noisy labels, its performance is affected by the adopted evaluation metric.

2. DATA SHAPLEY VALUE

This study implements the data SV in its data valuation approach for identifying training samples with noisy labels. SV [23] is a well-known measure in cooperative game theory for assigning a fair payoff to each player of the game by taking into account its contribution to all possible coalitions of the players. The training process of an ML/DL model can be assumed as a game where training data points are players that collaborate to achieve the highest model performance. Therefore, the SV can measure the importance of training samples in the performance of the ML/DL model [11–18]. Having a training set *D*, the SV of a data point $i \epsilon D$ is calculated as

$$SV_i = \frac{1}{N} \sum_{S \subseteq D-\{i\}} \frac{1}{\binom{N-1}{|S|}} [V\{S \cup i\} - V\{S\}], \tag{1}$$

where S is any possible subset of D not including i, N is the size of the training set D, and V is an evaluation metric for measuring the model performance. $V\{S\}$ is the performance of the model trained on the subset S and measured by the evaluation metric V. In other words, the SV of a training data point i is its average marginal contribution to all subsets of S, which can be interpreted as a quality measure for data assessment. Evaluating the precise SV, however, has exponential complexity in the number of players (data points). To compute the exact SV of each player, its marginal contribution to all possible coalitions of players needs to be computed, resulting in a computational complexity of $\mathcal{O}(2^N)$. Moreover, to evaluate SV for data points in an ML/DL task, calculating each marginal contribution, $V\{S\}$, requires training the model on S, which makes the computation process more expensive. To overcome this challenge, we use the permutation sampling method [11, 16, 18] resulting in

$$SV_i = \frac{1}{N!} \sum_{\pi \epsilon \Pi} [V\{S_i^\pi \cup i\} - V\{S_i^\pi\}], \tag{2}$$

where Π is all possible permutations of data points, $\pi \epsilon \Pi$ is a sample permutation of data points, and S_i^π is the set of training points coming before data point i in permutation π. Indeed, if we consider all possible joining orders of the data points, a coalition made up of N data points can form in $N!$ ways. Every new data point entering the coalition makes its marginal contribution to the previous data points in the coalition. Assuming that this joining process is a stochastic variable where all joining orders have the same likelihood of forming ($1/N!$), the SV of each data point can be considered as its expected marginal contribution. We approximate $\hat{SV}$ via the permutation sampling method, as the mean of m random samples, $SV_1, \ldots, SV_m$, drawn from the entire population of a data points' marginal contributions. In other words, we continue the permutation sampling until the average is empirically converged. In practice, the number of samples, m, required to reach the desired convergence is in order of $\mathcal{O}(N)$ and usually $3N$ sampling would be sufficient for the convergence of data $\hat{SV}$ [16]. It is worth mentioning that the SV of a data point is not a unique value and depends on the learning model, the evaluation metric, and other data points collaborating in the training game. So far, other studies have used a specific evaluation metric in their SV computation process and demonstrated the data SV capability in measuring the quality of data points. Here, we perform a set of experiments to investigate the effect of different evaluation metrics in measuring the SV, quantifying the quality of training points, and detecting noisy labels. We use the standard evaluation metrics in classification tasks, including accuracy, recall, and specificity, and discuss their effectiveness in detecting noisy labels in both classes of data. Our data valuation process, illustrated in Fig. 1, includes: i) training a deep classifier on a training set, ii) extracting high-resolution features from training CT images using the trained deep classifier and feeding them into a fast classifier, which is logistic regression in this study, iii) calculating the SV of each training point based on different evaluation metric, iv) analyzing the obtained SVs and detecting noisy labels.

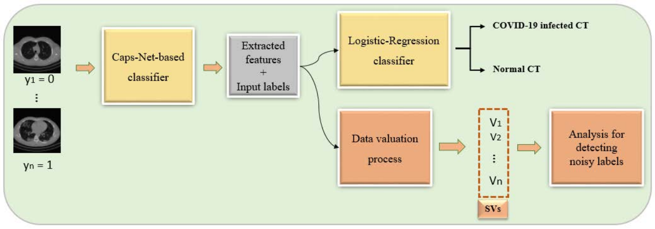
y1 = 0
⋮
yn = 1
Caps-Net-based classifier
Extracted features + Input labels
Logistic-Regression classifier
COVID-19 infected CT
Normal CT
Data valuation process
V1
V2
⋮
Vn
SVs
Analysis for detecting noisy labels

Fig. 1: The proposed framework to quantify the quality of data and discriminate COVID-19 infected CT images from normal ones.

The deep classifier used for discriminating COVID-19 infected CT images from the normal ones is a lightweight DL model containing two convolutional and two Capsule layers. A batch-normalization layer follows the first convolutional layer, and a max-pooling layer follows the second one. The ReLU activation function is applied after each convolutional layer to capture non-linear patterns. The number of channels for each convolutional layer is 64. The output of the max-pooling layer is reshaped and fed to a Capsule layer to extract high-resolution features from CT images. Finally, the last Capsule layer predicts the probability that each CT image belongs to infected or non-infected classes. It is noteworthy that screening COVID-19-infected CT images on a slice-level basis can be a primary step in detecting COVID-19 patients from healthy cases. We use a weighted loss function to tackle the imbalance dataset, considering a higher penalty to the samples from minority class, which is COVID-19 infected slices in our case. By extracting high-resolution features from the last Capsule layer, each CT image with a matrix size of 512×512 is presented in a 1×16 vector. Next, extracted features from the training set and their corresponding labels are used in the SV computation process. For more information about Capsule network-based classifiers and the weighted loss function, please refer to Reference [24].

3. DATASET DESCRIPTION

The dataset used in our experiments is a combination of two different open-access datasets. For non-COVID-19 CT images, we used a subset of 50 normal cases from our recently released dataset referred to as the "COVID-CT-MD" [25], which is available through Figshare [1]. For CT images with the evidence of COVID-19 lesions, we used a public dataset containing chest CT volumes of 10 COVID-19 patients [26], where three expert radiologists have annotated COVID-19 manifestations. In both datasets, the matrix size of the CT images is 512 × 512 pixels. We randomly split the dataset into three independent groups, as presented in table 1, including 100, 50, and 241 COVID-19 infected CT scans, and 400, 200, and 759 normal CT scans for training, validation, and test sets. The extracted features from the training set are used to train the logistic regression model over multiple permutation sampling, and the test set is used for measuring the model performance during the SV computation process.

[1] https://figshare.com/s/c20215f3d42c98f09ad0

Table 1: Training, validation and test sets used in our experiments.

	COVID-19	Normal	Total
Training set	100	400	500
Validation Set	50	200	250
Test Set	241	759	1000

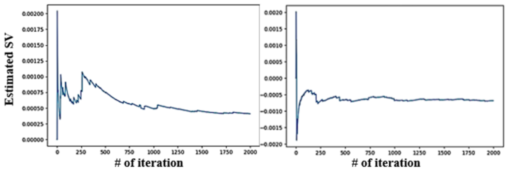

Fig. 2: The convergence of SV for randomly selected training samples. The x-axis indicates the number of iteration in SV computation process and the y-axis indicates the estimate of SV.

4. EXPERIMENTAL RESULTS

As the pre-processing step, each CT slice is normalized based on its pixel intensities' mean and standard deviation. Furthermore, we utilize the real-time data augmentation method to enhance the model's performance on unseen data. Each mini-batch of original images during the training process is transformed into synthetic images using conventional data augmentation strategies such as zooming, shifting, and shearing. Over the training' epochs, the model will observe each augmented image only once, resulting in an improvement in the model's generalization. In each mini-batch, 16 CT images are fed to the network. The Adam optimization algorithm with an initial learning rate of 0.001 minimizes the loss function over the training process. The number of passes through the training set is set to 100. However, to mitigate the model over-fitting, the training process will stop whenever the loss function on the validation set is not reduced over five epochs. Finally, the SV computation is performed using the training set as the players for training the logistic regression classifier and the test set for measuring the model performance trained on any possible subset of training points. As mentioned previously, we have 100 COVID-19 infected CT slices and 400 normal CT images as our training set. Since we aim to deal with noisy labels in the training set, we manually add some noises in the labels of our training set and investigate the capability of the SV obtained based on different evaluation metrics in detecting noisy labels. We run three experiments with three different noise levels, including 10%, 20%, and 30% in each class of data. Therefore, 10, 20, and 30 positive CT images and 40, 80, and 120 normal CT scans have incorrect labels in experiment I, II, and III, respectively.

In each experiment, first, we train the Capsule-network-based classifier using the training set with noisy labels. Next, the extracted features from the trained classifier are fed into a logistic regression classifier to compute the data SV. It is worth mentioning that computing the SV for each training point requires re-training the classifier on multiple subsets of training points, making the data evaluation process time-intensive. Hence, we use a fast classifier such as logistic regression to make the computation process affordable. We adopt different evaluation metrics, including accuracy, recall, and specificity, to calculate the data SV. We run the computation process for 2000 permutations to assure the convergence of the estimated data $\hat{SV}$s. Fig. 2 presents the convergence of the $\hat{SV}$ for two randomly selected training samples. As can be observed, the estimated $\hat{SV}$ converges after $3N = 1500$ permutations, which is in agreement with previous works [16].

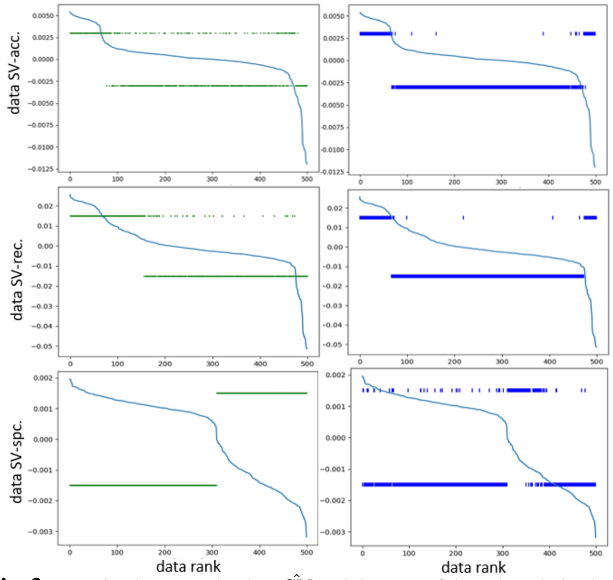

Fig. 3: Mapping between the data $\hat{SV}$ and the type of data class derived via different evaluation metrics. The green points and blue lines in top/bottom rows denote for noisy and ground-truth labels from positive/negative classes, respectively. (Plots are related to the Exp. III).

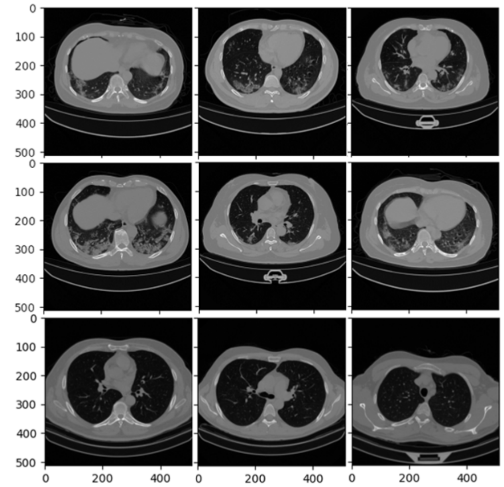

Fig. 4: Qualitative evaluation of CT images among the lowest data SVs. The accuracy, recall, and specificity metrics have been adopted for computing data SVs in the first, second, and third rows, respectively.

Next, we order the training points based on their estimated SV and mapped them with the corresponding label in our input labels that contains some noisy labels (green points in column left) and the ground-truth labels, which are correct labels (blue lines in right column), as visualized in Fig. 3. As can be observed, when adopting accuracy or recall metrics, all training points with the highest data SV belong to the positive class. In contrast, the specificity metric ranks all training points with the negative input labels before the data points with positive input labels. The plots also indicate that although the specificity metric puts all images with the positive input label among the lowest data SVs, the least valuable data points are the ones mislabeled as positive classes (their ground-truth label is negative). This reveals the capability of the SV in identifying noisy labels with the negative ground-truth labels (mislabeled as the positive images in the input label) when adopting the specificity metric. On the contrary, the majority of the least valuable training points obtained by the accuracy and recall metrics belong to negative input labels. However, according to the plots, the accuracy metric assigns more value to the images with negative input labels than recall metric. In addition, by comparing the plots in left and right columns, we can conclude that the least valuable data points obtained by both ac-

Table 2: Efficiency of data SV obtained based on different evaluation metrics in detecting noisy labels of each class of data in Exp I, II, and III with the noise level of 10%, 20%, and 30%.

Experiments		Positive CT images mislabeled as negative			Negative CT images mislabeled as positive		
		SV-acc	SV-rec	SV-spc	SV-acc	SV-rec	SV-spc
Exp. I (noise level: 10%)	Fraction of noisy labels in the lowest 10% of data SVs	1.00	1.00	0.00	0.28	0.28	0.68
	Fraction of noisy labels in the lowest 30% of data SVs	1.00	1.00	0.00	0.30	0.30	0.78
Exp. II (noise level: 20%)	Fraction of noisy labels in the lowest 20% of data SVs	1.00	1.00	0.00	0.23	0.18	0.96
	Fraction of noisy labels in the lowest 30% of data SVs	1.00	1.00	0.00	0.29	0.19	1.00
Exp. III (noise level: 30%)	Fraction of noisy labels in the lowest 30% of data SVs	0.97	0.97	0.00	0.35	0.07	1.00

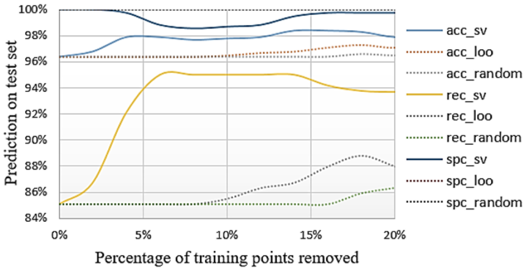

Fig. 5: Effects of removing least valuable data points on the performance.

curacy and recall metrics are the images mislabeled as the negative ones (have positive ground-truth labels). This indicates the possibility of detecting noisy labels with the positive ground-truth labels when using the accuracy or recall metrics in computing the SV. Furthermore, we investigate the effect of utilizing different evaluation metrics in the SV computation process for identifying noisy labels. We determined the fraction of noisy labels ranked as the 10%, 20%, and 30% of the lowest data SVs for experiments I, II, and III, respectively. We also consider the fraction of noisy labels ranked as the lowest 30% data SVs in all experiments. As demonstrated in Table 2, when adopting accuracy or recall metrics, all noisy labels with the positive ground-truth label (which had been incorrectly labeled as negative ones in input labels) have been correctly identified in experiments I and II. Conversely, when utilizing specificity metric in the SV computation process, no noisy label with the positive ground-truth label (mislabeled as negative images in the input labels) is ranked between the lowest 30% of data SVs. The results indicate that the adoption of the specificity metric has been more successful in detecting noisy labels with the negative ground-truth label (incorrectly labeled as positive ones in input labels). All noisy labels with the negative ground-truth labels (both correctly labeled and mislabeled ones) have been ranked in the lowest 30% data SVs.

Additionally, we have computed the leave-one-out (LOO) score for each data point, which is a common approach to assess data importance [27]. We eliminated 2% of the least valuable training points in each step, either based on their rank (on data SV and LOO score) or randomly. Then, a new logistic regression model is retrained. Fig. 5 represents the performance changes as data points are removed. It can be observed that drawing the least valuable data points identified by the data SV method improved the performance in terms of accuracy and recall more than using the LOO score or randomly removing data, which confirms capability of data SV in cleaning the training set. Fig. 4 demonstrates three random CT images ranked among the 30% lowest data SVs when adopting accuracy (first row), recall (second row), or specificity (third row) metrics for the computation of data SVs. As it can be observed, when using specificity, the normal CT images, incorrectly labeled as the infected class, have achieved the lowest data SVs. While when using the accuracy or recall, the positive CT images, incorrectly labeled as the negative ones, have obtained the lowest data SV. In particular, adopting specificity can detect noisy labels in the negative class, probably because it inherently emphasizes more on data points from the negative class. In contrast, since the recall puts more weight on the positive class, it is more capable of detecting noisy labels from the positive class. It is worth mentioning that accuracy that equally weighs data points from positive and negative classes, performs similar to recall in this task. One possible explanation is that the training set in our experiments is imbalanced with fewer data points from the positive class. As a result, data points from the positive class are of greater importance to maximizing the accuracy metric. It should be mentioned that, to the best of our knowledge, the adoption of different evaluation metrics in measuring the data SV has not been discussed in previous researches. However, Reference [17], which has performed an SV-based data valuation study using the accuracy metric on an X-ray imbalanced data set containing 1800 positive and 200 negative training points,illustrates that all the 100 training points with the highest SVs belong to the positive class. Besides, with the help of three radiologists, they figured out that out of the 100 lowest data SVs in their experiment, there were 65 mislabeled images where 80% of them were positive images that had been incorrectly labeled as negative ones. Indeed, their experiments are in accordance with our results, confirming that by utilizing the accuracy metric in computing data SV (for an imbalanced dataset where positive training points are in minority), the images with positive labels will receive the highest values. In addition, it would be more likely to detect mislabeled images with ground-truth positive labels (mislabeled as negative ones). Our findings show that, while the data SV has a lot of potential for detecting noisy labels in training sets, it is extremely dependent on the evaluation metric used.

5. CONCLUSION

The paper explored the effect of using different evaluation metrics in data SV computation and its capability and limitation in identifying noisy training labels. We examined standard evaluation metrics used for classification including accuracy, recall, and specificity in calculating the data SV on a dataset of chest CT scans. It is observed that while the data SV has potentials for detecting noisy training labels, this depends highly on the utilized evaluation metric. We also demonstrated that the data SV is not a unique value varying based on the incorporated evaluation metric. Different experiments have been conducted on a binary classification task with limited data. As a future direction, we aim to investigate capabilities and limitations of data SV based of different metrics for detecting noisy labels in a multi-institutional dataset and a multi-class classification problem. Exploring the dependency of the data SV on different types of learning algorithms is another future direction.

6. REFERENCES

[1] 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,” in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2097–2106, 2017.

[2] K. Zhang, X. Liu, J. Shen, Z. Li, Y. Sang, X. Wu, Y. Zha, W. Liang, C. Wang, K. Wang, *et al.*, “Clinically applicable ai system for accurate diagnosis, quantitative measurements, and prognosis of covid-19 pneumonia using computed tomography,” *Cell*, vol. 181, no. 6, pp. 1423–1433, 2020.

[3] A. Degerli, M. Ahishali, M. Yamac, S. Kiranyaz, M. E. Chowdhury, K. Hameed, T. Hamid, R. Mazhar, and M. Gabbouj, “Covid-19 infection map generation and detection from chest x-ray images,” *Health Information Science and Systems*, vol. 9, no. 1, pp. 1–16, 2021.

[4] G. Wang, X. Liu, C. Li, Z. Xu, J. Ruan, H. Zhu, T. Meng, K. Li, N. Huang, and S. Zhang, “A noise-robust framework for automatic segmentation of covid-19 pneumonia lesions from ct images,” *IEEE Transactions on Medical Imaging*, vol. 39, no. 8, pp. 2653–2663, 2020.

[5] L. Oakden-Rayner, “Exploring large-scale public medical image datasets,” *Academic radiology*, vol. 27, no. 1, pp. 106–112, 2020.

[6] X. Zhu and X. Wu, “Class noise vs. attribute noise: A quantitative study,” *Artificial intelligence review*, vol. 22, no. 3, pp. 177–210, 2004.

[7] S. Hu, Y. Gao, Z. Niu, Y. Jiang, L. Li, X. Xiao, M. Wang, E. F. Fang, W. Menpes-Smith, J. Xia, *et al.*, “Weakly supervised deep learning for covid-19 infection detection and classification from ct images,” *IEEE Access*, vol. 8, pp. 118869–118883, 2020.

[8] X. Wang, X. Deng, Q. Fu, Q. Zhou, J. Feng, H. Ma, W. Liu, and C. Zheng, “A weakly-supervised framework for covid-19 classification and lesion localization from chest ct,” *IEEE transactions on medical imaging*, vol. 39, no. 8, pp. 2615–2625, 2020.

[9] I. Laradji, P. Rodriguez, O. Manas, K. Lensink, M. Law, L. Kurzman, W. Parker, D. Vazquez, and D. Nowrouzezahrai, “A weakly supervised consistency-based learning method for covid-19 segmentation in ct images,” in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 2453–2462, 2021.

[10] C. Xue, Q. Dou, X. Shi, H. Chen, and P.-A. Heng, “Robust learning at noisy labeled medical images: Applied to skin lesion classification,” in *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*, pp. 1280–1283, IEEE, 2019.

[11] S. Maleki, L. Tran-Thanh, G. Hines, T. Rahwan, and A. Rogers, “Bounding the estimation error of sampling-based shapley value approximation,” *arXiv preprint arXiv:1306.4265*, 2013.

[12] S. M. Lundberg and S.-I. Lee, “A unified approach to interpreting model predictions,” in *Proceedings of the 31st international conference on neural information processing systems*, pp. 4768–4777, 2017.

[13] L. Merrick and A. Taly, “The explanation game: Explaining machine learning models using shapley values,” in *International Cross-Domain Conference for Machine Learning and Knowledge Extraction*, pp. 17–38, Springer, 2020.

[14] E. Song, B. L. Nelson, and J. Staum, “Shapley effects for global sensitivity analysis: Theory and computation,” *SIAM/ASA Journal on Uncertainty Quantification*, vol. 4, no. 1, pp. 1060–1083, 2016.

[15] E. Štrumbelj and I. Kononenko, “Explaining prediction models and individual predictions with feature contributions,” *Knowledge and information systems*, vol. 41, no. 3, pp. 647–665, 2014.

[16] A. Ghorbani and J. Zou, “Data shapley: Equitable valuation of data for machine learning,” in *International Conference on Machine Learning*, pp. 2242–2251, PMLR, 2019.

[17] S. Tang, A. Ghorbani, R. Yamashita, S. Rehman, J. A. Dunnmon, J. Zou, and D. L. Rubin, “Data valuation for medical imaging using shapley value and application to a large-scale chest x-ray dataset,” *Scientific reports*, vol. 11, no. 1, pp. 1–9, 2021.

[18] R. Jia, D. Dao, B. Wang, F. A. Hubis, N. Hynes, N. M. Gürel, B. Li, C. Zhang, D. Song, and C. J. Spanos, “Towards efficient data valuation based on the shapley value,” in *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 1167–1176, PMLR, 2019.

[19] T. Wang, J. Rausch, C. Zhang, R. Jia, and D. Song, “A principled approach to data valuation for federated learning,” in *Federated Learning*, pp. 153–167, Springer, 2020.

[20] N. Khuri and S. Parsons, “A value-based approach for training of classifiers with high-throughput small molecule screening data,” in *Proceedings of the 12th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics*, pp. 1–10, 2021.

[21] G. Hinton, S. Sabour, and N. Frosst, “Matrix capsules with EM routing,” *6th International Conference on Learning Representations, ICLR 2018 - Conference Track Proceedings*, pp. 1–29, 2018.

[22] A. Mohammadi, Y. Wang, N. Enshaei, P. Afshar, F. Naderkhani, A. Oikonomou, J. Rafiee, H. Rodrigues de Oliveira, S. Yanushkevich, and K. N. Plataniotis, “Diagnosis/Prognosis of COVID-19 Chest Images via Machine Learning and Hypersignal Processing: Challenges, opportunities, and applications,” *IEEE Signal Processing Magazine*, vol. 38, pp. 37–66, sep 2021.

[23] L. S. Shapley, *17. A value for n-person games*. Princeton University Press, 2016.

[24] S. Heidarian, P. Afshar, N. Enshaei, F. Naderkhani, M. J. Rafiee, F. B. Fard, K. Samimi, S. F. Atashzar, A. Oikonomou, K. N. Plataniotis, *et al.*, “Covid-fact: A fully-automated capsule network-based framework for identification of covid-19 cases from chest ct scans,” *Frontiers in Artificial Intelligence*, vol. 4, 2021.

[25] P. Afshar, S. Heidarian, N. Enshaei, F. Naderkhani, M. J. Rafiee, A. Oikonomou, F. Babaki Fard, K. Samimi, K. Plataniotis, and A. Mohammadi, “COVID-CT-MD: COVID-19 Computed Tomography (CT) Scan Dataset Applicable in Machine Learning and Deep Learning,” 2020.

[26] J. Ma, Y. Wang, X. An, C. Ge, Z. Yu, J. Chen, Q. Zhu, G. Dong, J. He, Z. He, *et al.*, “Towards efficient covid-19 ct annotation: A benchmark for lung and infection segmentation,” *arXiv e-prints*, pp. arXiv–2004, 2020.

[27] R. D. Cook, “Detection of influential observation in linear regression,” *Technometrics*, vol. 42, no. 1, pp. 65–68, 2000.