This MICCAI paper is the Open Access version, provided by the MICCAI Society. It is identical to the accepted version, except for the format and this watermark; the final published version is available on SpringerLink.

Modeling and Understanding Uncertainty in Medical Image Classification

Aobo Chen, Yangyi Li, Wei Qian, Kathryn Morse, Chenglin Miao, and Mengdi Huai

Iowa State University, Ames IA 50011, USA
{aobochen,liyangyi,wqi,kamorse,cmiao,mdhuai}@iastate.edu

Abstract. Medical image classification is an important task in many different medical applications. The past years have witnessed the success of Deep Neural Networks (DNNs) in medical image classification. However, traditional softmax outputs produced by DNNs fail to estimate uncertainty in medical image predictions. Contrasting with conventional uncertainty estimation approaches, conformal prediction (CP) stands out as a model-agnostic and distribution-free methodology that constructs statistically rigorous uncertainty sets for model predictions. However, existing exact full conformal methods involve retraining the underlying DNN model for each test instance with each possible label, demanding substantial computational resources. Additionally, existing works fail to uncover the root causes of medical prediction uncertainty, making it difficult for doctors to interpret the estimated uncertainties associated with medical diagnoses. To address these challenges, in this paper, we first propose an efficient approximate full CP method, which involves tracking the gradient updates contributed by these samples during training. Subsequently, we design an interpretation method that uses these updates to identify the top-k most influential training samples that significantly impact models' uncertainties. Extensive experiments on real-world medical image datasets are conducted to verify the effectiveness of the proposed methods.

Keywords: Medical image classification · Deep learning · Uncertainty estimation · Model explanations.

1 Introduction

Effectively classifying medical images is crucial for aiding clinical care and treatment [29, 21]. For example, analysis X-ray is the best approach for diagnosing pneumonia [21], responsible for approximately 50,000 deaths annually in the US. However, classifying pneumonia from chest X-rays needs professional radiologists, who are often rare and expensive resources in many regions. Traditional machine learning methods, such as SVMs [17], have been used in medical image classification for quite a long time. However, their performance is far from the practical standard, and their development has quite slowed in recent years. Also, the feature extraction and selection are time-consuming and vary across different

objects [13]. Recently, deep learning [15, 24] has significantly advanced medical image classification [28, 5, 14], showing substantial performance gains across different types of medical images, including CT/MRI and ultrasound images.

In medical image classification, accurately estimating prediction uncertainty is vital for reducing diagnostic errors. While DNNs naturally produce softmax outputs, they lack a solid theoretical uncertainty guarantee. Along with the significance of uncertainty estimation, a paradigm called *conformal prediction* (CP) [23, 7, 4, 16] has spawned, which is a simple yet powerful paradigm for creating statistically rigorous uncertainty sets for pre-trained networks. Critically, the sets are valid in a distribution-free sense without distribution and model assumptions. Conformal prediction requires a user-specified significance level to restrict the frequency of errors that the model is allowed to make. For example, a significance level of 0.1 means that the model makes at most 10% erroneous predictions. For skin-lesion classification [25, 19], this means that predictions are not a single label, but instead a set (e.g., {"melanocytic nevus", "melanoma"}), which covers the true label with $1 - 10\% = 90\%$ probability on average. It matches the intuition of clinical decision-making by providing a set of possible labels that rule in or rule out certain diseases similar to a differential diagnosis.

Existing works on CP can be divided into: split CP [20, 11] and full CP [7, 23]. Compared with split CP, full CP achieves stronger validity guarantees and typically smaller prediction set sizes by ensuring that conformal sets are calibrated based on the full data distribution. This is achieved at a significant computational cost, as full CP requires the model to be retrained for each test data against every possible label to assess uncertainty accurately. This is impractical for DNNs due to resource demands. While [18] introduces the influence functions-based approximation method, recent literature [2] reveals the limitations and fragility of influence functions in DNNs. The inefficiency challenge is pronounced in scenarios requiring rapid real-time prediction uncertainties.

Additionally, existing research fails to elucidate the origins of prediction uncertainties, a gap particularly critical in the medical domain. In healthcare, accurately identifying the causes of prediction uncertainty is beneficial for enhancing patient outcomes and ensuring diagnostic accuracy. Identifying which training samples contribute most significantly to the model's uncertainty can provide valuable insights into the model's behavior [22, 30]. This insight is vital for medical practitioners and researchers aiming to refine the model, whether by enriching the dataset with more diverse medical image samples, fine-tuning the model's parameters to better capture the nuances of complex medical conditions, or employing advanced training strategies to bolster the model's predictive accuracy and reliability. Therefore, the ability to explain prediction uncertainties becomes paramount, especially in medical applications.

To address the aforementioned challenges, in this paper, we first propose TAFCP, a training **T**rajectory-based **A**pproximate **F**ull **CP** method for medical image classification. Specifically, in our method, we expand the training trajectory of the pre-trained model with a Taylor series, and then formalize the approximation of the deletion and addition of specific samples from the pre-

trained model via a single-step, closed-form update. In this way, our method can effectively alleviate the high computational demands of full CP by eliminating the need for model retraining. In addition, we also develop a novel **Un**certainty **Ex**planation method (UnEX) to identify the top-k most influential training samples impacting the model's uncertainty. This insight enhances understanding of prediction uncertainty, helping developers create transparent and interpretable diagnostic uncertainty tools for clinicians. Extensive experiments are conducted across a variety of medical image classification tasks to demonstrate the effectiveness and potential of our proposed methods for practical medical applications.

2 Methodology

In this section, we introduce our approximate full CP method in medical image classification, which efficiently constructs conformal sets without retraining. Following this, we present a novel interpretation method that offers the transparency necessary for clinicians to understand prediction uncertainties effectively.

Without loss of generality, in this paper, we consider the medical image classification tasks. Let Y denote the number of classes. We denote the available training dataset as $D = \{z_i = (x_i, y_i)\}_{i=1}^N$, where $x_i \in \mathbb{R}^d$ is d-dimensional, N denotes the number of training samples and $y_i \in [Y]$. Given D, we can train a classifier $f(\cdot; W^*) : \mathbb{R}^d \to \mathbb{R}^M$ by solving the following optimization

$$W^* = \arg\min_{W \in \mathcal{W}} \sum_{i=1}^{N} \ell((x_i, y_i); W), \tag{1}$$

where ℓ is defined based on the cross-entropy loss (the de-facto choice for classification). Many established optimization schemes are derived from mini-batch stochastic gradient descent (SGD) [6, 24]. Formally, mini-batch SGD can be described as $W_{t+1} = W_t - \eta \frac{\partial \ell}{\partial W}|_{W_t, \bar{x}_t}$, where weights at step t are obtained using the weights from step $t-1$, $\bar{x}_t$ is the mini-batch data used at step t, and η is the learning rate. For the well-trained classifier $f(\cdot; W^*)$, it outputs class probabilities of the incoming patient x^{pat} (i.e., $f(x^{pat}; W^*) = [f_1(x^{pat}; W^*), \cdots, f_Y(x^{pat}; W^*)]$). The predicted label for x^{pat} is $y(x^{pat}; W^*) = \arg\max_{\hat{y} \in [Y]} f_{\hat{y}}(x^{pat}; W^*)$.

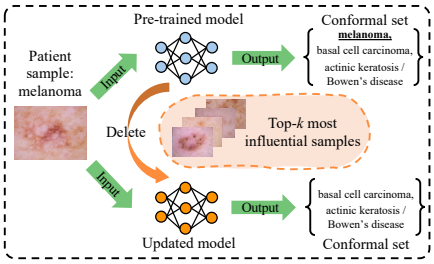
Training trajectory-based approximate full CP method. Note that when using traditional full CP to construct the conformal prediction set $\mathcal{C}_\varepsilon(x^{pat})$ of possible labels for patient x^{pat}, as indicated in Algorithm 1 in Fig. 1a, statistical test is conducted for each possible label $\hat{y} \in [Y]$ to decide if it should be included in $\mathcal{C}_\varepsilon(x^{pat})$. Importantly, the statistical test requires computing a non-conformity score α_i by retraining the model for each sample in the newly augmented training set $D \cup \{(x^{pat}, \hat{y})\}$. Then, a p-value is computed, and a decision is taken based on the threshold ε. Assuming that the training data D and x^{pat} are exchangeable [23], the true label $y^*(x^{pat})$ of patient x^{pat} satisfies

$$P(y^*(x^{pat}) \in \mathcal{C}_\varepsilon(x^{pat})) \geq 1 - \varepsilon. \tag{2}$$

Algorithm 1 Full CP

for x^{pat} among the incoming patients **do**
  for $\hat{y} \in [Y]$ **do**
    $\hat{z} = (x^{pat}, \hat{y})$
    for $z_i \in D \cup \{\hat{z}\}$ **do**
      if traditional full CP (retraining) **then**
        $\tilde{W}_{D\cup\{\hat{z}\}\setminus\{z_i\}} = \text{argmin}_{W\in\mathcal{W}} \sum_{\tilde{z}\in D\cup\{\hat{z}\}\setminus\{z_i\}} \ell(\tilde{z}; W)$
      else if approximate full CP (TAFCP) **then**
        $\tilde{W}_{D\cup\{\hat{z}\}\setminus\{z_i\}} = W_D - \frac{\eta E}{b}(\nabla\ell(\hat{z}; W_D) - \nabla\ell(z_i; W_D))$
      $\alpha_i = \ell(z_i; \tilde{W}_{D\cup\{\hat{z}\}\setminus\{z_i\}})$
    $p_{(x^{pat},\hat{y})} = \frac{\#\{i=1,\ldots,N+1 : \alpha_i \geq \alpha_{N+1}\}}{N+1}$
    Include $\hat{y}$ to $\mathcal{C}_\varepsilon(x^{pat})$ if $p_{(x^{pat},\hat{y})} > \varepsilon$
return $\mathcal{C}_\varepsilon(x^{pat})$ for each patient x^{pat}

(a) Full CP



(b) Uncertainty explanations

Fig. 1. An overview of our approximate full CP and uncertainty explanations.

Notably, the data exchangeability assumption is much weaker than the i.i.d assumption [23]. However, as illustrated in Fig. 1a, calculating the conformal set requires retraining the underlying DNN model with the added data and then executing a leave-one-out retraining procedure. Yet, such retraining for each test case and label is resource-heavy and impractical for large networks.

The challenge of running traditional full CP is the computation of the nonconformity scores $\alpha_i = \ell(z_i, \tilde{W}_{D\cup\{\hat{z}\}\setminus\{z_i\}})$. Each score is determined by computing the loss of the model at sample $z_i \in D\cup\{\hat{z}\}$ after adding sample $\hat{z}$ and removing sample z_i from the model's training data D. To avoid retraining each time, we propose to approximate the contribution of adding and removing the samples to the model, and then evaluate its loss at z_i. To understand the impact of a target sample $x^* \in D$ on the final weights W^*, we apply Taylor series expansion to the sequence of SGD (Stochastic Gradient Descent) updates [6, 31, 12]. We begin with the definition of a single SGD learning update: $W_1 = W_0 - \eta\frac{\partial\ell}{\partial W}|_{W_0,\bar{x}_0}$, where W_0 denotes the weights at step 0 and x_0 denotes the data sampled at step 0. Here, we make no constraints on what W_0 is. Then, we can obtain W_2 (the weight obtained at step 2) as $W_2 = W_0 - \eta\frac{\partial\ell}{\partial W}|_{W_0,\bar{x}_0} - \eta\frac{\partial\ell}{\partial W}|_{W_1,\bar{x}_1}$. Based on that $W_1 = W_0 - \eta\frac{\partial\ell}{\partial W}|_{W_0,\bar{x}_0}$, we can rewrite W_2 as

$$W_2 = W_0 - \eta\frac{\partial\ell}{\partial W}|_{W_0,\bar{x}_0} - \eta\frac{\partial\ell}{\partial W}|_{W_0-\eta\frac{\partial\ell}{\partial W}|_{W_0,\bar{x}_0},\bar{x}_1} \tag{3}$$

$$\approx W_0 - \eta(\frac{\partial\ell}{\partial W}|_{W_0,\bar{x}_0} + \frac{\partial\ell}{\partial W}|_{W_0,\bar{x}_1} + \frac{\partial^2\ell}{\partial^2 W}|_{W_0,\bar{x}_1}(-\eta\frac{\partial\ell}{\partial W}|_{W_0,\bar{x}_0})). \tag{4}$$

To scale the above to t sequential updates, note that the hessian terms from expansion are recursively interdependent. Thus, we have the below approximation

$$W_t \approx W_0 - \eta\sum_{i=1}^{t-1}\frac{\partial\ell}{\partial W}|_{W_0,\bar{x}_i} + \sum_{i=1}^{t-1} g(i), \tag{5}$$

where $g(i)$ is recursively defined as $g(i) = -\eta\frac{\partial^2\ell}{\partial^2 W}|_{W_0,\bar{x}_i}(-\eta\sum_{j=0}^{i-1}\frac{\partial\ell}{\partial W}|_{W_0,\bar{x}_j} + \sum_{j=0}^{i-1} g(i))$ with $g(0) = 0$. In Eq. (5), the terms in the first sum are simply gradients taken with respect to the initial model weights W_0, and $\bar{x}_i$ follows

the order we give data to the DNN model. Note that the exact order does not matter as we simply add them. Therefore, the effect of the target sample $x^* \in D$ (provided at any step in training) on this first sum is a model gradient computed with respect to W_0 and x^*. Since the second term is a negligible error for practical applications [12], we can ignore this term. Based on the above, to unlearn sample x^* from the pre-trained model W^*, we perform the below update

$$\tilde{W} \leftarrow W^* + \frac{\eta E}{b} \nabla \ell(x^*; W^*), \tag{6}$$

where b is the batch size and E is the number of epochs. In the above, we simply add back these gradients by adding $\frac{\eta E}{b} \frac{\partial \ell}{\partial W}|_{W_0, x^*}$ to the final weights to unlearn data without retraining. In a similar way, to add a new data $x^* \notin D$, we need to calculate $\tilde{W} \leftarrow W^* - \frac{\eta E}{b} \nabla \ell(x^*; W^*)$, as shown in line 8 of Algorithm 1.

Based on the above, for patient x^{pat}, we can efficiently obtain its α_i (in line 9 of Algorithm 1 in Fig. 1a) without retraining. Notably, our method gives a substantial speed-up over full CP to construct conformal sets by requiring only a single update for sample modifications, enhancing scalability for large datasets.

Uncertainty explanations. In medical applications, interpreting uncertainties associated with diagnosis predictions is essential. Our goal here is to identify the top-k most influential training samples for the generated conformal sets. For simplicity, we illustrate the main idea of our proposed UnEx via an example in Fig. 1b, where we focus on the top-k most influential training samples, the absence of which would lead to excluding a label $y^{tar} \in [Y]$ (e.g., "melanoma") from the original conformal set $\mathcal{C}_\varepsilon(x^{pat})$ (e.g., {"melanoma", "basal cell carcinoma", "actinic keratosis / Bowen's disease"}). For each $z_i = (x_i, y_i) \in D$, we define a discrete indication parameter $\xi_i \in \{0, 1\}$ to indicate whether z_i should be the top-k most influential ($\xi_i = 1$) or not ($\xi_i = 0$). The influential dataset $\tilde{D}_k$ is denoted as $\tilde{D}_k = D \circ \Phi = \{z_i | z_i \in D \text{ and } \xi_i = 1\}$, where $\Phi = \{\xi_i\}_{i=1}^N$. To select the most influential subset $\tilde{D}_k$, we formulate the following optimization

$$\begin{aligned} \max_{\tilde{D}_k = D \circ \Phi} \mathbb{I}[p(x^{pat}, y^{tar}) < \varepsilon] + \sum_{y \in \mathcal{C}_\varepsilon(x^{pat}) \setminus y^{tar}} \mathbb{I}[p(x^{pat}, y) > \varepsilon] \\ \text{s.t. } \tilde{W} \leftarrow W^* + \sum_{z_i \in D} \xi_i \cdot \frac{\eta E}{b} \nabla \ell(z_i; W^*), \end{aligned} \tag{7}$$

where $\Phi = \{\xi_i \in \{0, 1\}\}_{i=1}^N$, and $p(x^{pat}, y)$ is defined in line 10 in Fig. 1a. The removal of the optimized subset $\tilde{D}_k$ would result in the deletion of y^{tar} from the original conformal set, i.e., $y^{tar} \notin \tilde{\mathcal{C}}_\varepsilon(x^{pat})$ and $y^{tar} \in \mathcal{C}_\varepsilon(x^{pat})$, where $\tilde{\mathcal{C}}_\varepsilon(x^{pat})$ is derived based on $D \setminus \tilde{D}_k$. Thus, the above first loss is designed to enforce the exclusion of y^{tar} from $\tilde{\mathcal{C}}_\varepsilon(x^{pat})$, while the second one aims to guarantee the inclusion of the remaining labels. Additionally, for the above constraint, we adopt our proposed approximate update method in Eq. (6) to provide a closed-form update for the deletion of the selected training samples. In this way, we can circumvent the extensive computational and storage requirements associated

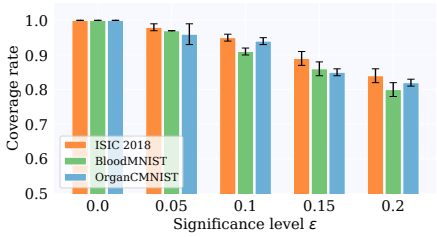


Fig. 2. Validity of TAFCP at varying significance levels.

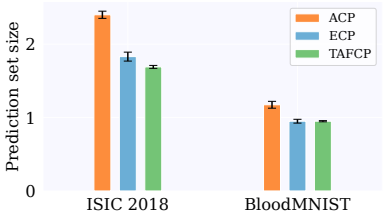


Fig. 3. Efficiency comparison. A smaller size implies better efficiency.

with such bi-level optimization problems [27, 10], which typically necessitate the full retraining to solve the inner problem before updating the outer one. Notably, we can easily extend the explanation framework in Eq. (7) to generate different uncertainty explanations, such as elucidating the exclusion of the true label.

However, directly solving this framework is infeasible due to the discrete nature of the above indication function-based loss, which involves the combinational search across the training samples to identify the influential subset $\tilde{D}_k$ and is computationally infeasible. To address these challenges, we propose an efficient empirical search method, which begins with an empty set and gradually adds the most influential sample. The search process continues until adding a sample causes the targeted label to shift from being inside to outside the generated conformal set. *Note that the optimal value of k varies across different incoming patients, reflecting the personalized aspect of our method.*

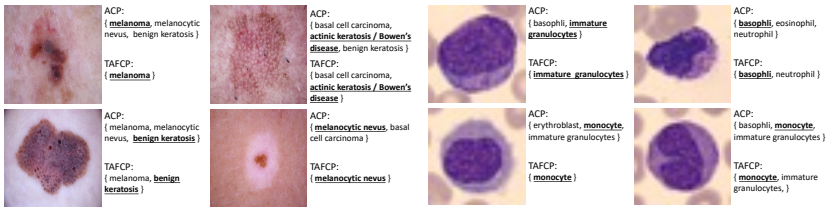
3 Experiments

In this section, we perform comprehensive experiments to validate the effectiveness of our methods. *Due to space limitations, more experimental details and results can be found in the supplementary material.*

3.1 Experimental Setup

Datasets. In experiments, we adopt the following real-world medical image datasets: ISIC 2018 [8], BloodMNIST [1], and OrganCMNIST [3, 26]. ISIC 2018 consists of 10,015 dermoscopic images spanning 7 skin diseases. BloodMNIST is a dataset of normal peripheral blood, featuring 17,092 images of individual cells. OrganCMNIST comprises 23,583 2D computed tomography (CT) images extracted from the central slices of 3D bounding boxes in the coronal view.

Baselines. In experiments, we adopt the *exact full CP* (ECP) with full retraining [23, 7] and *influence functions-based approximate full CP* (ACP) [18] as the baselines. Specifically, ACP employs influence functions (second-order) to approximately update the model when constructing the conformal sets.



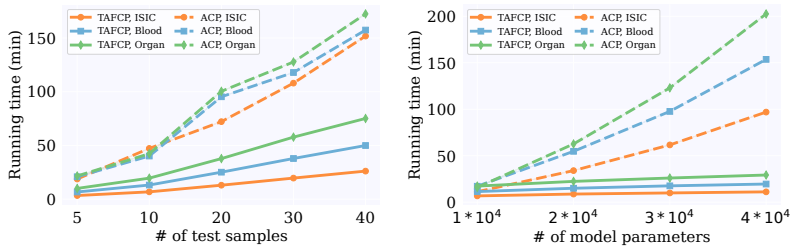
(a) ISIC 2018 (b) BloodMNIST

Fig. 4. Conformal sets comparison. Bold and underlined phrases mean true labels.

Implementation details. For the adopted datasets, we employ ResNet-18 [9] and a 2-layer convolutional network (CNN). The CNN incorporates convolutional blocks with 16 and 32 features, respectively, and integrates various linear layers. We train the model for 40 epochs, using the SGD optimizer with a learning rate of 1e–3 and a batch size of 100. All experiments are conducted across 5 trials, and we report the average results.

3.2 Experimental Results

Validity. In Fig. 2, we investigate the validity of our proposed TAFCP on ISIC 2018, BloodMNIST, and OrganCMNIST datasets. We report the empirical coverage rate (the percentage of true labels that are held in the conformal sets) at varying significance levels. We observe that the coverage rates achieved by TAFCP are at least $1 - \varepsilon$, which consistently satisfies the desired probability expectations in Eq. (2). For example, TAFCP attains a coverage rate of 0.8 on the BloodMNIST dataset with a significance level set to 0.2. Therefore, TAFCP demonstrates a satisfying empirical coverage rate across various datasets, indicating the utility and effectiveness of our proposed approach.



(a) Data complexity (b) Model complexity

Fig. 5. Running time comparison on data complexity and model complexity.

Efficiency. In Fig. 3, we explore the efficiency of our proposed TAFCP. We present the average set size at a significance level of 0.1 across different medical image datasets. As depicted, TAFCP consistently surpasses the ACP baseline, and is comparable to the exact full CP with retraining. For instance, on the ISIC 2018 dataset, the average set size of TAFCP is approximately 1.69, compared to about 2.4 for ACP. We also provide examples of the conformal sets in Fig. 4. We observe that both TAFCP and ACP correctly output true labels, while TAFCP produces smaller sets than ACP. This indicates that TAFCP outperforms ACP in efficiently modeling uncertainty within medical image classification tasks.

Running time. In Fig. 5, we illustrate the running time of TAFCP. We compare it with the ACP baseline under various data complexity and model complexity within three medical image classification tasks. It is observed that TAFCP requires significantly less running time than ACP. For instance, in Fig. 5a, when testing 40 test samples on the ISIC 2018 dataset, TAFCP completes in about 26 minutes, while influence functions-based ACP requires over 150 minutes, where there is approximately a fivefold difference. Moreover, Fig. 5b reveals that as the model complexity increases, the running time of ACP exhibits polynomial growth, whereas the running time of TAFCP shows linear behavior, significantly reducing the computation cost. These results show TAFCP's superior computational efficiency in addressing the time complexity associated with full CP.

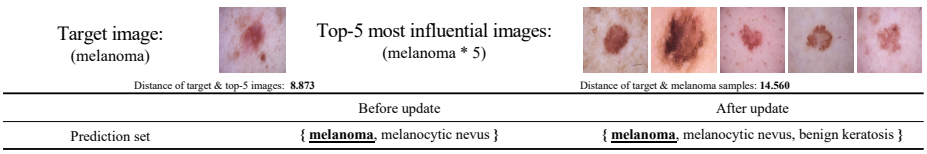


Fig. 6. Visualization results of our proposed UnEX.

Uncertainty explanations. Here, we provide a visualization of the generated uncertainty explanations on the ISIC 2018 dataset with ResNet-18. In Fig. 6, we visualize the identified top-5 most influential training samples for the target image (i.e., "melanoma") on the left side. Initially, the original prediction set of this target image is {"melanoma", "melanocytic nevus"} and these influential samples are labeled as "melanoma". After deleting these images from the training set, the prediction set expands and includes the "benign keratosis" label. Note that a larger conformal set size indicates greater uncertainty in the model's predictions. This increased model uncertainty is due to the close proximity of these excluded images (labeled as "melanoma") to the target sample (classified as "melanoma") that results in a lack of training in this critical area. This gap in training amplifies the uncertainty in the model's predictions. Therefore, the "benign keratosis" label is included in the conformal set for this target sample.

4 Conclusion

In this research, we first present a training trajectory-based approximate full CP method, which can efficiently estimate prediction uncertainties with significantly reduced computational complexity through a single-step, closed-form update. Following this, we then develop an uncertainty interpretation method that uses these closed-form gradient updates to identify the top-k most influential training samples affecting the model's uncertainty levels. Identifying these critical samples allows medical imaging experts to focus on targeted enhancements, thereby reducing uncertainty. Extensive experiments on real-world medical image datasets demonstrate the practicality and efficiency of our methods.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Acevedo, A., Merino, A., Alférez, S., Molina, Á., Boldú, L., Rodellar, J.: A dataset of microscopic peripheral blood cell images for development of automatic recognition systems. Data in brief **30** (2020)
2. Basu, S., Pope, P., Feizi, S.: Influence functions in deep learning are fragile. In: International Conference on Learning Representations (2020)
3. Bilic, P., Christ, P., Li, H.B., Vorontsov, E., Ben-Cohen, A., Kaissis, G., Szeskin, A., Jacobs, C., Mamani, G.E.H., Chartrand, G., et al.: The liver tumor segmentation benchmark (lits). Medical Image Analysis **84**, 102680 (2023)
4. Boström, H., Johansson, U., Vesterberg, A.: Predicting with confidence from survival data. In: Conformal and Probabilistic Prediction and Applications. pp. 123–141. PMLR (2019)
5. Carneiro, G., Nascimento, J., Bradley, A.P.: Unregistered multiview mammogram analysis with pre-trained deep learning models. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 652–660. Springer (2015)
6. Chen, B., Xu, Y., Shrivastava, A.: Fast and accurate stochastic gradient estimation. Advances in Neural Information Processing Systems **32** (2019)
7. Cherubin, G., Chatzikokolakis, K., Jaggi, M.: Exact optimization of conformal predictors via incremental and decremental learning. In: International Conference on Machine Learning. pp. 1836–1845. PMLR (2021)
8. Codella, N., Rotemberg, V., Tschandl, P., Celebi, M.E., Dusza, S., Gutman, D., Helba, B., Kalloo, A., Liopyris, K., Marchetti, M., et al.: Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (isic). arXiv preprint arXiv:1902.03368 (2019)
9. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
10. Huang, W.R., Geiping, J., Fowl, L., Taylor, G., Goldstein, T.: Metapoison: Practical general-purpose clean-label data poisoning. Advances in Neural Information Processing Systems **33**, 12080–12091 (2020)
11. Humbert, P., Le Bars, B., Bellet, A., Arlot, S.: One-shot federated conformal prediction. In: International Conference on Machine Learning. pp. 14153–14177. PMLR (2023)

12. Jastrzębski, S., Kenton, Z., Arpit, D., Ballas, N., Fischer, A., Bengio, Y., Storkey, A.: Three factors influencing minima in sgd. arXiv preprint arXiv:1711.04623 (2017)
13. Kermany, D.S., Goldbaum, M., Cai, W., Valentim, C.C., Liang, H., Baxter, S.L., McKeown, A., Yang, G., Wu, X., Yan, F., et al.: Identifying medical diagnoses and treatable diseases by image-based deep learning. cell **172**(5), 1122–1131 (2018)
14. Korot, E., Guan, Z., Ferraz, D., Wagner, S.K., Zhang, G., Liu, X., Faes, L., Pontikos, N., Finlayson, S.G., Khalid, H., et al.: Code-free deep learning for multimodality medical image classification. Nature Machine Intelligence **3**(4), 288–298 (2021)
15. LeCun, Y., Bengio, Y., Hinton, G.: Deep learning. nature **521**(7553), 436–444 (2015)
16. Li, Y., Chen, A., Qian, W., Zhao, C., Lidder, D., Huai, M.: Data poisoning attacks against conformal prediction. In: Forty-first International Conference on Machine Learning
17. Maglogiannis, I.G., Zafiropoulos, E.P.: Characterization of digital medical images utilizing support vector machines. BMC Medical Informatics and Decision Making **4**(1), 1–9 (2004)
18. Martinez, J.A., Bhatt, U., Weller, A., Cherubin, G.: Approximating full conformal prediction at scale via influence functions. In: Proceedings of the AAAI Conference on Artificial Intelligence. pp. 6631–6639 (2023)
19. Mehrtens, H., Bucher, T., Brinker, T.J.: Pitfalls of conformal predictions for medical image classification. In: International Workshop on Uncertainty for Safe Utilization of Machine Learning in Medical Imaging. pp. 198–207. Springer (2023)
20. Olsson, H., Kartasalo, K., Mulliqi, N., Capuccini, M., Ruusuvuori, P., Samaratunga, H., Delahunt, B., Lindskog, C., Janssen, E.A., Blilie, A., et al.: Estimating diagnostic uncertainty in artificial intelligence assisted pathology using conformal prediction. Nature communications **13**(1), 7761 (2022)
21. Organization, W.H., et al.: Standardization of interpretation of chest radiographs for the diagnosis of pneumonia in children. Tech. rep., World Health Organization (2001)
22. Qian, W., Zhao, C., Li, Y., Ma, F., Zhang, C., Huai, M.: Towards modeling uncertainties of self-explaining neural networks via conformal prediction. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 14651–14659 (2024)
23. Shafer, G., Vovk, V.: A tutorial on conformal prediction. Journal of Machine Learning Research **9**(3) (2008)
24. Thudi, A., Deza, G., Chandrasekaran, V., Papernot, N.: Unrolling sgd: Understanding factors influencing machine unlearning. In: 2022 IEEE 7th European Symposium on Security and Privacy (EuroS&P). pp. 303–319. IEEE (2022)
25. Tschandl, P., Rosendahl, C., Kittler, H.: The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. Scientific data **5**(1), 1–9 (2018)
26. Xu, X., Zhou, F., Liu, B., Fu, D., Bai, X.: Efficient multiple organ localization in ct image using 3d region proposal network. IEEE transactions on medical imaging **38**(8), 1885–1898 (2019)
27. Zhang, H., Gao, J., Su, L.: Data poisoning attacks against outcome interpretations of predictive models. In: Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining. pp. 2165–2173 (2021)
28. Zhang, Y., Chen, H., Wei, Y., Zhao, P., Cao, J., Fan, X., Lou, X., Liu, H., Hou, J., Han, X., et al.: From whole slide imaging to microscopy: Deep microscopy

adaptation network for histopathology cancer image classification. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 360–368. Springer (2019)
29. Zhang, Y., Gao, J., Zhou, M., Wang, X., Qiao, Y., Zhang, S., Wang, D.: Text-guided foundation model adaptation for pathological image classification. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 272–282. Springer (2023)
30. Zhao, C., Qian, W., Shi, Y., Huai, M., Liu, N.: Automated natural language explanation of deep visual neurons with large models. arXiv preprint arXiv:2310.10708 (2023)
31. Zheng, S., Meng, Q., Wang, T., Chen, W., Yu, N., Ma, Z.M., Liu, T.Y.: Asynchronous stochastic gradient descent with delay compensation. In: International Conference on Machine Learning. pp. 4120–4129. PMLR (2017)