

---

# Deep AUC Maximization for Medical Image Classification: Challenges and Opportunities

---

Tianbao Yang

Department of Computer Science  
The University of Iowa, Iowa City, IA 52242  
tianbao-yang@uiowa.edu

## Abstract

In this extended abstract, we will present and discuss opportunities and challenges brought about by a new deep learning method by AUC maximization (aka **Deep AUC Maximization** or **DAM**) for medical image classification. Since AUC (aka area under ROC curve) is a standard performance measure for medical image classification, hence directly optimizing AUC could achieve a better performance for learning a deep neural network than minimizing a traditional loss function (e.g., cross-entropy loss). Recently, there emerges a trend of using deep AUC maximization for large-scale medical image classification. In this paper, we will discuss these recent results by highlighting (i) the advancements brought by stochastic non-convex optimization algorithms for DAM; (ii) the promising results on various medical image classification problems. Then, we will discuss challenges and opportunities of DAM for medical image classification from three perspectives, feature learning, large-scale optimization, and learning trustworthy AI models.

## 1 A Brief history of AUC Maximization

AUC maximization has a history of almost two decades. During these two decades, there have been a variety of methods for optimizing the AUC score. In order to better position deep AUC maximization, we start by introducing the formulations and objectives for AUC maximization.

**Definitions and Formulations.** Let  $(\mathbf{x}, \mathbf{y}) \sim \mathcal{P}$  denote an input data and label pair following a distribution  $\mathcal{P}$ , where  $\mathbf{x} \in \mathbb{R}^d$  and  $\mathbf{y} \in \{1, -1\}$ . Let  $h_{\mathbf{w}}(\cdot) : \mathbb{R}^d \rightarrow \mathbb{R}$  denote a predictive model (e.g., a deep neural network, a linear model). The ROC curve is obtained by plotting the true positive rate (TPR) vs. the false positive rate (FPR) by varying the prediction threshold. AUC can be calculated as the Riemann integral of the function TPR vs FPR. However, the complex nature of computing Riemann integral makes it difficult to design optimization algorithms towards maximizing AUC for learning the model  $h_{\mathbf{w}}(\cdot)$ . The most popular approach is to use the probability interpretation of AUC [25], i.e.,  $\text{AUC}(h_{\mathbf{w}}(\cdot), \mathcal{P}) = \Pr(h_{\mathbf{w}}(\mathbf{x}) > h_{\mathbf{w}}(\mathbf{x}') | y = 1, y' = -1)$ , which means that AUC (on the population level) is equivalent to the probability that a randomly selected positive data is ranked higher than a randomly selected negative data by the predictive function, where  $(\mathbf{x}, y = 1)$  is a random positive data and  $(\mathbf{x}', y' = -1)$  is a random negative data. On a given dataset  $\mathcal{D} = \mathcal{D}_+ \cup \mathcal{D}_-$ , where  $\mathcal{D}_+ = \{(\mathbf{x}_i, y_i) \in \mathcal{D} : y_i = 1\}$  and  $\mathcal{D}_- = \{(\mathbf{x}_i, y_i) \in \mathcal{D} : y_i = -1\}$ , the empirical AUC score can be computed as

$$\text{AUC}(h_{\mathbf{w}}(\cdot), \mathcal{D}) = \frac{1}{n_+} \frac{1}{n_-} \sum_{\mathbf{x}_i \in \mathcal{D}_+} \sum_{\mathbf{x}'_i \in \mathcal{D}_-} s(h_{\mathbf{w}}(\mathbf{x}_i) - h_{\mathbf{w}}(\mathbf{x}'_i)), \quad (1)$$

where  $n_+ = |\mathcal{D}_+|$ ,  $n_- = |\mathcal{D}_-|$ ,  $s(a) = 1$  if  $a > 0$ ,  $s(a) = 1/2$  if  $a = 0$ , and  $s(a) = 0$  if  $a < 0$ .

**Full Batch based Methods (2000 - 2010).** Earlier works for AUC maximization use full batch based methods, which process all examples at each iteration. To the best of our knowledge, the

earliest work dates back to 2000 [28], which considers the ordinal regression problem and proposes support vector machine (SVM) formulation in the sense of pairwise classification. Later, the AUC maximization problem has been tackled by [55] using gradient-based methods, by [20] for learning decision-tree models, by [21] in the framework of boosting, by [31] in the framework of structured SVM. Some works also proposed speed-up techniques by reducing the number of pairs [29, 47] or by reducing the number of iterations [60]. Nevertheless, these full batch based algorithms could suffer a quadratic time complexity in the worst-case or a super-linear (e.g. log-linear) time complexity per-iteration, which makes them not scalable to large datasets.

**Online Methods (2010 - 2015).** To the best of our knowledge, Zhao et al. [62] is the first work that considers AUC maximizing in the online learning fashion. To deal with large-scale data, they proposed to maintain a dynamic buffer to store some historical data for updating the model parameter. This online learning approach was also studied in several later works [33, 34]. In [33], the authors also proposed mini-batch stochastic gradient methods that update the model parameters based on all data pairs in the mini-batch. However, since these online methods do not consider all data pairs their error bound or regret bound depends on the size of the buffer or mini-batch (e.g.,  $1/\sqrt{B}$ ). Hence, these algorithms will not converge unless the buffer size of mini-batch size  $B$  is infinitely large. Gao et al. [22] proposed an one-pass AUC maximization algorithm based on the pairwise square loss for learning a linear model, which optimizes an online version of AUC at each iteration that pairs each received data with all historical data. To avoid storing all historical data, they leverage the structure of the square loss and maintain and update mean and covariance matrix of data.

**Stochastic mini-batch based Methods (2016 - present).** [57] is a milestone work for stochastic optimization of AUC. They restricted their attention to the pairwise square loss and proposed to transform the non-decomposable objective into a decomposable min-max optimization problem, which favors stochastic methods based on mini-batch of data without explicitly constructing the pairs. Later on, the convergence of stochastic optimization of AUC based on the min-max formulation was improved in [37, 41, 42]. The min-max formulation also serves as the basis for most recent works on DAM discussed in next section.

## 2 DAM and Applications in Medical Image Classification

Recently, there is a trend of DAM based on large-scale stochastic optimization algorithms for solving the non-convex min-max formulation of the pairwise square loss or its variants. We briefly discuss the developments for large-scale non-convex min-max optimization and its application to AUC maximization and then present some successful empirical studies on medical image classification.

**Non-Convex Min-Max Optimization and Stochastic DAM.** Stochastic non-convex min-max optimization algorithms were first analyzed in [46] with provable convergence guarantee. Thereafter, the theoretical developments have been the major topic of a wave of studies [7, 24, 30, 35, 38, 54, 56]. Liu et al. [36] developed the first practical and provable stochastic algorithms for DAM based on the min-max formulation of the pairwise square loss function, which enjoy a fast convergence rate. Recently, DAM has been also studied in the framework of federated learning [23, 58].

**DAM for Medical Image Classification.** To the best of our knowledge, [59] is the first work that evaluates the performance of DAM on large-scale medical image data with hundreds of thousands images, which is two orders larger than that was used in earlier works, e.g., [53]. They proposed a new objective for robust AUC maximization to alleviate the issues of the square loss, namely the sensitivity to noisy data and the adverse effect on easy data. The new loss function can be also cast into a min-max objective, to which all existing non-convex min-max optimization algorithms can be applied. It was shown to be more robust than the commonly used square loss, while enjoying the same advantage in terms of large-scale stochastic optimization. They conducted extensive empirical studies of DAM on four difficult medical image classification tasks, including (i) classification of chest X-ray images for identifying many threatening diseases, (ii) classification of images of skin lesions for identifying melanoma, (iii) classification of mammogram for breast cancer screening, and (iv) classification of microscopic images for identifying tumor tissue. In Table 1, we summarize their results on various medical image datasets. The percentage in the column "Improvements" shows the improvement over the baseline method that is trained by minimizing the standard cross-entropy loss, the column "Competition Results" shows their rank over all participating teams in competitions. The authors also released a library for DAM called LibAUC <sup>1</sup>.

---

<sup>1</sup>[www.libauc.org](http://www.libauc.org)

Table 1: Summary of DAM’s performance on several medical image classification tasks from [59]. The CheXpert data was released in the Stanford CheXpert competition [1], the Melanoam data was released in the 2020 Kaggle Melanoma Competition [2].

Dataset	Image Domain	#pos/#all	# Training	Improvements	Competition Results
CheXpert	Chest X-ray	20.21%	224,316	2%	1/150+
Melanoma	Skin Lesion	7.1%	46,131	1%	33/3314
DDSM+	Mammogram	13%	55,000	1.5%	NA
PatchCamelyon	Microscopic	1%	148,960	5%	NA

### 3 Challenges and Opportunities

In this section, we will discuss outstanding challenges and new opportunities of DAM for medical image classification.

**Feature Learning.** Feature learning is an important capability of deep learning for tackling unstructured image data. The current practice of DAM uses a two-stage approach: the first stage is to learn the encoder network by optimizing the traditional cross-entropy loss and the second stage is to fine tune the encoder network and to learn the classifier by DAM [59]. It is still not fully understood why optimizing the AUC loss in an end-to-end fashion does not yield better feature representations, and it remains an open problem how to learn better encoder networks by using DAM. One direction is to improve the end-to-end learning paradigm that could enjoy the benefit of both feature learning and robust classifier learning. Another direction is to consider self-supervised pre-training methods on large-scale unlabeled medical datasets. This approach was recently explored in [52, 61, 4]. But its success on downstream tasks of using DAM remains to be demonstrated. In this direction, we could consider pre-training on much larger medical datasets than those used in existing studies and demonstrate the performance of DAM on multiple downstream medical image classification tasks.

**Large-scale Optimization.** Although large-scale optimization algorithms for DAM have been developed, there are still many open problems to be addressed. Below, we will list several important questions. (i) How is the performance of optimizing a min-max loss compared with that of optimizing the conventional pairwise surrogate loss based on mini-batch data? It remains unclear which approach is faster and more robust for deep learning. (ii) How to optimize partial AUC for deep learning? Partial AUC maximization is much more challenging than standard AUC maximization since the former involves the ordering of prediction scores among a large number of examples. (iii) How to optimize area under precision-recall curve (AUPRC)? AUPRC is shown to be more appropriate for assessing the performance of a classifier on highly imbalanced data [18, 49]. Its close sibling performance metric named averaged precision (AP) and its optimization has attracted much attention in information retrieval and computer vision [9–11, 14–16, 19, 27, 40, 43, 45, 48]. Recently, there is a breakthrough on large-scale AP optimization with provable convergence guarantee by Qi et al. [44]. Nevertheless, it is still an open area for developing faster and robust methods for deep AUPRC maximization.

**Learning fair and interpretable AI models.** Building trustworthy AI is important for healthcare domains, in particular medical image classification. Two issues are of foremost importance, namely fairness [5, 6, 12, 39] and interpretability [13, 26, 3]. Although these issues have received tremendous attention in the literature for medical image classification [17, 51, 50], developing fair and interpretable DAM methods remains to be explored. Some outstanding questions and work include (i) how to develop scalable in-processing algorithms for optimizing AUC under AUC-based fairness constraints [8, 32]; (ii) how to develop scalable and interpretable DAM methods; (iii) evaluating these fairness-aware and interpretable AUC optimization methods on large-scale medical image datasets.

### 4 Conclusions

In this extended abstract, we have discussed the history of AUC maximization in the last two decades. We then presented the recent studies on non-convex deep AUC maximization and its applications on various medical image classification problems. Finally, we discussed some outstanding challenges and new opportunities of deep AUC maximization, which serve as good research topics in the next few years.

## References

- [1] Chexpert competition. <https://stanfordmlgroup.github.io/competitions/chexpert/>.
- [2] Kaggle competition: Siim-istic melanoma classification. <https://www.kaggle.com/c/siim-istic-melanoma-classification>.
- [3] Sercan O Arik and Tomas Pfister. Protoattend: Attention-based prototypical learning. *Journal of Machine Learning Research*, 21:1–35, 2020.
- [4] Shekoofeh Azizi, Basil Mustafa, Fiona Ryan, Zachary Beaver, Jan Freyberg, Jonathan Deaton, Aaron Loh, Alan Karthikesalingam, Simon Kornblith, Ting Chen, Vivek Natarajan, and Mohammad Norouzi. Big self-supervised models advance medical image classification, 2021.
- [5] Solon Barocas, Moritz Hardt, and Arvind Narayanan. *Fairness and Machine Learning*. fairml-book.org, 2019. <http://www.fairmlbook.org>.
- [6] Rachel K. E. Bellamy, Kuntal Dey, Michael Hind, Samuel C. Hoffman, Stephanie Houde, Kalapriya Kannan, Pranay Lohia, Jacquelyn Martino, Sameep Mehta, Aleksandra Mojsilovic, Seema Nagar, Karthikeyan Natesan Ramamurthy, John T. Richards, Diptikalyan Saha, Prasanna Sattigeri, Moninder Singh, Kush R. Varshney, and Yunfeng Zhang. AI fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias. *CoRR*, abs/1810.01943, 2018.
- [7] Radu Ioan Boct and Axel Böhm. Alternating proximal-gradient steps for (stochastic) nonconvex-concave minimax problems. *arXiv preprint arXiv:2007.13605*, 2020.
- [8] Daniel Borkan, Lucas Dixon, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. Nuanced metrics for measuring unintended bias with real data for text classification. In *Companion proceedings of the 2019 world wide web conference*, pages 491–500, 2019.
- [9] Andrew Brown, Weidi Xie, Vicky Kalogeiton, and Andrew Zisserman. Smooth-ap: Smoothing the path towards large-scale image retrieval. In *European Conference on Computer Vision*, pages 677–694. Springer, 2020.
- [10] Christopher Burges, Robert Ragno, and Quoc Le. Learning to rank with nonsmooth cost functions. In B. Schölkopf, J. Platt, and T. Hoffman, editors, *Advances in Neural Information Processing Systems*, volume 19. MIT Press, 2007.
- [11] Fatih Cakir, Kun He, Xide Xia, Brian Kulis, and Stan Sclaroff. Deep metric learning to rank. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.
- [12] Simon Caton and Christian Haas. Fairness in machine learning: A survey. *CoRR*, abs/2010.04053, 2020.
- [13] Chaofan Chen, Oscar Li, Chaofan Tao, Alina Jade Barnett, Jonathan Su, and Cynthia Rudin. This looks like that: deep learning for interpretable image recognition. *arXiv preprint arXiv:1806.10574*, 2018.
- [14] Kean Chen, Jianguo Li, Weiyao Lin, John See, Ji Wang, Lingyu Duan, Zhibo Chen, Changwei He, and Junni Zou. Towards accurate one-stage object detection with ap-loss. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.
- [15] Kean Chen, Weiyao Lin, John See, Ji Wang, Junni Zou, et al. Ap-loss for accurate one-stage object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.
- [16] Wei Chen, Tie-Yan Liu, Yanyan Lan, Zhiming Ma, and Hang Li. Ranking measures and loss functions in learning to rank. In *Proceedings of the 22nd International Conference on Neural Information Processing Systems, NIPS’09*, page 315–323, Red Hook, NY, USA, 2009. Curran Associates Inc.
- [17] Valeriia Cherepanova, Vedant Nanda, Micah Goldblum, John P. Dickerson, and Tom Goldstein. Technical challenges for training fair neural networks. *CoRR*, abs/2102.06764, 2021.

- [18] Jesse Davis and Mark Goadrich. The Relationship Between Precision-Recall and ROC Curves. In *ICML '06: Proceedings of the 23rd international conference on Machine learning*, pages 233–240, New York, NY, USA, 2006. ACM.
- [19] Elad Eban, Mariano Schain, Alan Mackey, Ariel Gordon, Rif A. Saurous, and Gal Elidan. Scalable learning of non-decomposable objectives. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2017.
- [20] C Ferri, PA Flach, and J Hernández-Orallo. Learning decision trees using the area under the roc curve. In Claude Sammut and Achim Hoffmann, editors, *Proceedings of the 19th International Conference on Machine Learning*, pages 139 – 146. Morgan Kaufmann, 2002.
- [21] Yoav Freund, Raj Iyer, Robert E. Schapire, and Yoram Singer. An efficient boosting algorithm for combining preferences. *J. Mach. Learn. Res.*, 4(null):933–969, December 2003.
- [22] Wei Gao, Rong Jin, Shenghuo Zhu, and Zhi-Hua Zhou. One-pass auc optimization. In *International conference on machine learning*, pages 906–914, 2013.
- [23] Zhishuai Guo, Mingrui Liu, Zhuoning Yuan, Li Shen, Wei Liu, and Tianbao Yang. Communication-efficient distributed stochastic AUC maximization with deep neural networks. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 3864–3874. PMLR, 2020.
- [24] Zhishuai Guo, Zhuoning Yuan, Yan Yan, and Tianbao Yang. Fast objective and duality gap convergence for non-convex strongly-concave min-max problems. *CoRR*, abs/2006.06889, 2020.
- [25] J.A. Hanley and Barbara Mcneil. The meaning and use of the area under a receiver operating characteristic (roc) curve. *Radiology*, 143:29–36, 05 1982.
- [26] Peter Hase, Chaofan Chen, Oscar Li, and Cynthia Rudin. Interpretable image recognition with hierarchical prototypes. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 7, pages 32–40, 2019.
- [27] Paul Henderson and Vittorio Ferrari. End-to-end training of object class detectors for mean average precision. In *Computer Vision – ACCV 2016*, pages 198–213. Springer International Publishing, 2017.
- [28] Ralf Herbrich, Thore Graepel, and Klaus Obermayer. Large Margin Rank Boundaries for Ordinal Regression. In *Advances in Large Margin Classifiers*, chapter 7, pages 115–132. The MIT Press, 1999.
- [29] Alan Herschtal and Bhavani Raskutti. Optimising area under the ROC curve using gradient descent. In Carla E. Brodley, editor, *Machine Learning, Proceedings of the Twenty-first International Conference (ICML 2004), Banff, Alberta, Canada, July 4-8, 2004*, volume 69 of *ACM International Conference Proceeding Series*. ACM, 2004.
- [30] Feihu Huang, Shangqian Gao, Jian Pei, and Heng Huang. Accelerated zeroth-order momentum methods from mini to minimax optimization. *arXiv preprint arXiv:2008.08170*, 2020.
- [31] Thorsten Joachims. A support vector method for multivariate performance measures. In *Proceedings of the 22nd International Conference on Machine Learning, ICML '05*, page 377–384, New York, NY, USA, 2005. Association for Computing Machinery.
- [32] Nathan Kallus and Angela Zhou. The fairness of risk scores beyond classification: Bipartite ranking and the xAuc metric. *Advances in neural information processing systems*, 32:3438–3448, 2019.
- [33] Purushottam Kar, Harikrishna Narasimhan, and Prateek Jain. Online and stochastic gradient methods for non-decomposable loss functions. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 1, NIPS'14*, page 694–702, Cambridge, MA, USA, 2014. MIT Press.

- [34] Purushottam Kar, Bharath Sriperumbudur, Prateek Jain, and Harish Karnick. On the generalization ability of online learning algorithms for pairwise loss functions. In Sanjoy Dasgupta and David McAllester, editors, *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pages 441–449, Atlanta, Georgia, USA, 17–19 Jun 2013. PMLR.
- [35] Tianyi Lin, Chi Jin, and Michael I. Jordan. On gradient descent ascent for nonconvex-concave minimax problems. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, pages 6083–6093, 2020.
- [36] Mingrui Liu, Zhuoning Yuan, Yiming Ying, and Tianbao Yang. Stochastic AUC maximization with deep neural networks. In *8th International Conference on Learning Representations (ICLR)*, 2020.
- [37] Mingrui Liu, Xiaoxuan Zhang, Zaiyi Chen, Xiaoyu Wang, and Tianbao Yang. Fast stochastic auc maximization with  $o(1/n)$ -convergence rate. In *International Conference on Machine Learning*, pages 3189–3197, 2018.
- [38] Luo Luo, Haishan Ye, Zhichao Huang, and Tong Zhang. Stochastic recursive gradient descent ascent for stochastic nonconvex-strongly-concave minimax problems. In *Advances in Neural Information Processing Systems 33 (NeurIPS)*, 2020.
- [39] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *CoRR*, abs/1908.09635, 2019.
- [40] P. Mohapatra, Michal Rolinek, C. V. Jawahar, V. Kolmogorov, and M. Kumar. Efficient optimization for rank-based loss functions. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3693–3701, 2018.
- [41] Michael Natole, Yiming Ying, and Siwei Lyu. Stochastic proximal algorithms for auc maximization. In *International Conference on Machine Learning*, pages 3710–3719, 2018.
- [42] Michael Natole, Yiming Ying, and Siwei Lyu. Stochastic auc optimization algorithms with linear convergence. *Frontiers in Applied Mathematics and Statistics*, 5, 2019.
- [43] Kemal Oksuz, Baris Can Cam, Emre Akbas, and Sinan Kalkan. A ranking-based, balanced loss function unifying classification and localisation in object detection. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 15534–15545. Curran Associates, Inc., 2020.
- [44] Qi Qi, Youzhi Luo, Zhao Xu, Shuiwang Ji, and Tianbao Yang. Stochastic optimization of area under precision-recall curve for deep learning with provable convergence. *CoRR*, abs/2104.08736, 2021.
- [45] Tao Qin, Tie-Yan Liu, and Hang Li. A general approximation framework for direct optimization of information retrieval measures. Technical Report MSR-TR-2008-164, November 2008.
- [46] Hassan Rafique, Mingrui Liu, Qihang Lin, and Tianbao Yang. Non-convex min-max optimization: Provable algorithms and applications in machine learning. *Optimization Methods and Software*, 2020.
- [47] Alain Rakotomamonjy. Support vector machines and area under roc curves. Technical report, 2004.
- [48] Michal Rolinek, Vit Musil, Anselm Paulus, Marin Vlastelica, Claudio Michaelis, and Georg Martius. Optimizing rank-based metrics with blackbox differentiation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.
- [49] Takaya Saito and Marc Rehmsmeier. The precision-recall plot is more informative than the roc plot when evaluating binary classifiers on imbalanced datasets. *PLOS ONE*, 10(3):1–21, 03 2015.
- [50] Kathryn Schutte, Olivier Moindrot, Paul Hérent, Jean-Baptiste Schiratti, and Simon Jégou. Using stylegan for visual interpretability of deep learning models on medical images, 2021.

- [51] Laleh Seyyed-Kalantari, Guanxiong Liu, Matthew B. A. McDermott, and Marzyeh Ghassemi. Chexclusion: Fairness gaps in deep chest x-ray classifiers. *CoRR*, abs/2003.00827, 2020.
- [52] Hari Sowrirajan, Jingbo Yang, Andrew Y. Ng, and Pranav Rajpurkar. Moco pretraining improves representation and transferability of chest x-ray models. *CoRR*, abs/2010.05352, 2020.
- [53] Jeremias Sulam, Rami Ben-Ari, and Pavel Kisilev. Maximizing auc with deep learning for classification of imbalanced mammogram datasets. In *VCBM*, pages 131–135, 2017.
- [54] Quoc Tran-Dinh, Deyi Liu, and Lam M. Nguyen. Hybrid variance-reduced SGD algorithms for minimax problems with nonconvex-linear function. In *Advances in Neural Information Processing Systems 33 (NeurIPS)*, 2020.
- [55] Lian Yan, Robert Dodier, Michael C. Mozer, and Richard Wolniewicz. Optimizing classifier performance via an approximation to the wilcoxon-mann-whitney statistic. In *Proceedings of the Twentieth International Conference on International Conference on Machine Learning, ICML'03*, page 848–855. AAAI Press, 2003.
- [56] Yan Yan, Yi Xu, Qihang Lin, Wei Liu, and Tianbao Yang. Optimal epoch stochastic gradient descent ascent methods for min-max optimization. In *Advances in Neural Information Processing Systems 33 (NeurIPS)*, 2020.
- [57] Yiming Ying, Longyin Wen, and Siwei Lyu. Stochastic online auc maximization. In *Advances in neural information processing systems*, pages 451–459, 2016.
- [58] Zhuoning Yuan, Zhishuai Guo, Yi Xu, Yiming Ying, and Tianbao Yang. Federated deep AUC maximization for heterogeneous data with a constant communication complexity. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 12219–12229. PMLR, 2021.
- [59] Zhuoning Yuan, Yan Yan, Milan Sonka, and Tianbao Yang. Robust deep AUC maximization: A new surrogate loss and empirical studies on medical image classification. In *International Conference on Computer Vision*, volume abs/2012.03173, 2020.
- [60] Xinhua Zhang, Ankan Saha, and S. V. N. Vishwanathan. Smoothing multivariate performance measures. *J. Mach. Learn. Res.*, 13(1):3623–3680, December 2012.
- [61] Yuhao Zhang, Hang Jiang, Yasuhide Miura, Christopher D. Manning, and Curtis P. Langlotz. Contrastive learning of medical visual representations from paired images and text. *CoRR*, abs/2010.00747, 2020.
- [62] Peilin Zhao, Steven C. H. Hoi, Rong Jin, and Tianbao Yang. Online auc maximization. In *ICML*, pages 233–240, 2011.