arXiv:2107.08964v1 [cs.CV] 19 Jul 2021

Transductive image segmentation: Self-training and effect of uncertainty estimation

Konstantinos Kamnitsas[1], Stefan Winzeck[1], Evgenios N. Kornaropoulos[2,5], Daniel Whitehouse[2], Cameron Englman[2], Poe Phyu[3], Norman Pao[4], David K. Menon[2], Daniel Rueckert[1,6], Tilak Das[3], Virginia F.J. Newcombe[2], and Ben Glocker[1]

[1] Department of Computing, Imperial College London, UK
[2] Division of Anaesthesia, Department of Medicine, University of Cambridge, UK
[3] Dept Radiology, Cambridge Univ. Hospitals NHS Foundation Trust, UK
[4] Dept Emergency Medicine, Cambridge Univ. Hospitals NHS Foundation Trust, UK
[5] Clinical Sciences, Diagnostic Radiology, Lund University, Sweden
[6] Klinikum rechts der Isar, Technical University of Munich, Germany

Abstract. Semi-supervised learning (SSL) uses unlabeled data during training to learn better models. Previous studies on SSL for medical image segmentation focused mostly on improving model generalization to unseen data. In some applications, however, our primary interest is not generalization but to obtain optimal predictions on a specific unlabeled database that is fully available during model development. Examples include population studies for extracting imaging phenotypes. This work investigates an often overlooked aspect of SSL, transduction. It focuses on the quality of predictions made on the unlabeled data of interest when they are included for optimization during training, rather than improving generalization. We focus on the self-training framework and explore its potential for transduction. We analyze it through the lens of Information Gain and reveal that learning benefits from the use of calibrated or under-confident models. Our extensive experiments on a large MRI database for multi-class segmentation of traumatic brain lesions shows promising results when comparing transductive with inductive predictions. We believe this study will inspire further research on transductive learning, a well-suited paradigm for medical image analysis.

1 Introduction

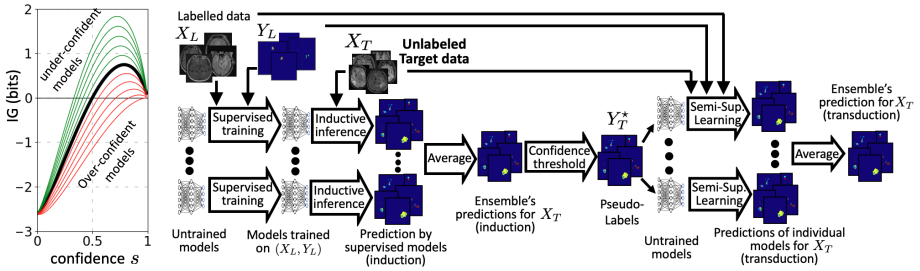
The predominant paradigm for developing ML models is supervised training using labeled data. Because labels are scarce in many applications, these models often suffer from generalization issues on real-world heterogeneous data. To alleviate this, semi-supervised learning (SSL) [5], aims to extract additional useful information from unlabeled data provided during training. Most commonly, the trained model is afterwards applied to *new, previously unseen* unlabeled data at test time, predicts by *induction*, and its generalization is measured by the accuracy on this new data. Model predictions made for the unlabeled data during

training, as part of the optimization, are commonly discarded. For certain applications, however, the objective is to obtain predictions on a particular unlabeled database which may be *fully available at time of model development*, rather than generalization to other data afterwards. Applications of this type include population analysis on retrospective data to extract image phenotypes (e.g. size or location of pathology). In such a setting, it is desirable to *directly* optimize predictions for the specific unlabeled data, which is the concept of *transduction*. In this context, this study takes a fresh look on SSL and evaluates the potential of transduction for image segmentation, which has been previously under-explored.

Transductive learning is a form of SSL. Formally, SSL approaches assume two databases are available during model development: a *labeled* database $D_L = \{X_L \cup Y_L\} = \{(x_{L,i}, y_{L,i})\}_{i=1}^{N}$ and an *unlabeled* database $D_U = \{X_U\} = \{x_{U,i}\}_{i=1}^{M}$, with $y_{L,i} \in \mathcal{Y} = \{1, ..., C\}$ and C the number of classes. Common SSL approaches construct model $f_\theta(x) = p(y|x;\theta) \in \mathbb{R}^C$ and learn parameters θ such that it will approximate distribution $q(y|x)$ that generates the labels. This is commonly done by jointly minimizing a supervised cost $J_s(X_L, Y_L, f_\theta)$ (e.g. cross entropy) and an unsupervised cost, $J_u(X_U, \theta)$. For SSL focusing on *induction*, the goal is to obtain optimized model f'_θ which *generalizes* to new, unseen and unlabeled *target* (or *test*) data D_T. In contrast, in SSL focusing on *transduction*, it is assumed that the unlabeled *target* data is available during training (i.e. $D_U = D_T$). The goal is then to obtain the best possible predictions $Y'_T = \{y'_{T,i}\}_{i=1}^{M}$ that approximate the unknown true labels $Y_T = \{y_{T,i}\}_i^M$. Some methods, such as those employed in this study, can be used both for induction and transduction. However, in transduction the predictions on D_T are directly optimized during training, and may thus be better than predictions via induction, as the latter is subject to degradation due to distributional shift and the generalization gap [5,22].

Research on transductive SSL spans decades, with prime example the transductive SVM [22]. In medical imaging, transduction has been mostly used for learning over graphs [23,7] and few-shot learning [2]. Related work also explored transductive learning of a meta-learner for combining predictions from multiple models [25]. We are instead interested in the common setting where a relation graph between samples does not exist, significant amounts of labels are available (not few-shot) and we investigate whether and how can transduction reliably benefit standard segmentation models. In these settings, most prior work on SSL developed inductive methods and evaluated generalization to unseen data (e.g. [1,6,24,15,12]). Benefits by SSL were primarily shown when labeled data are very limited (few tens of images) and diminish with more labels [1,6,24,15]. In practical settings where more labeled data are available, the predominant paradigm is still supervised learning (e.g. large scale studies). Therefore, SSL methods that reliably offer improvements in such practical settings are still desirable.

This study makes the following contributions: (a) We present a theoretical analysis of the SSL framework of self-training via the lenses of Information Gain and reveal that learning benefits from well-calibrated or under-confident predictive uncertainty. This motivates us to introduce model ensembling within the framework, which is currently the most reliable method for obtaining cali-



(a) (b)

Fig. 1: a) Expected IG when a sample predicted with confidence s is assigned a 1-hot pseudo-label can be computed as $IG = \delta s H - (1 - \delta s)H$, where H the entropy in the predicted class-posteriors. Lines represent IG for different values of δ. $\delta = 1$ models perfect calibrated s, $0 < \delta < 1$ over-confident and $\delta > 1$ under-confident. For illustration, here H was calculated for 7-class problem, with s probability of most probable class and assuming $(1-s)/6$ for remaining classes.b) Self-training with ensembles for improved confidence calibration and learning. Predictions for target data are obtained via transduction in the end of 2nd training phase.

brated uncertainty. (b) We assess quality of predictions obtained via transductive SSL and show consistent improvements over induction with our proposed framework. (c) We perform extensive evaluation on a large multi-center database for the challenging task of multi-class segmentation of traumatic brain injuries (TBI), including a blinded assessment via manual segmentation refinement, and demonstrate that transductive SSL with our framework can provide consistent improvements over induction, even when hundreds of labeled data are used. We believe this study will motivate further research into transductive learning, which is a suitable, but currently overlooked, paradigm for medical image analysis.

2 Methodology

We analyse and improve *self-training* [21], a popular framework for inductive SSL with neural networks [17,1]. Hence its transductive potential is of interest.

2.1 Transductive learning via self-training

In this framework, model f_θ (e.g. a neural net) is first trained via supervised cost $J_s(X_L, Y_L, \theta)$ to obtain θ'. Then, $f_{\theta'}$ is applied to unlabeled data, which in our transductive setting are *target* data X_T. The predicted labels $Y'_T = \arg\max_c(f_{\theta'}(X_T))$ with highest class-posterior are then used as *pseudo-labels* $Y_T^\star$ added to the training set. If the model provides confidence estimate s_i for i-th sample, a *confidence threshold* t can be chosen, and only confident predictions $(s_i > t)$ are used as pseudo-labels (in segmentation, uncertain pixels are masked out from the loss). The model is then re-trained on the extended training set

minimizing $J_{total,ps} = J_s(X_L, Y_L, \theta) + \beta J_s(X_T, Y_T^\star, \theta)$, with β a hyper-parameter. This gives new parameters θ''. Inductive SSL strives to learn model $f_{\theta''}$ that will *generalize* better than $f_{\theta'}$ on *new* data. Instead, transductive learning focuses on predictions $Y_T'' = f_{\theta''}(X_T)$ to improve them over Y_T' obtained via standard supervision. We now analyse via Information Gain the framework, which reveals how uncertainty calibration of $f_{\theta'}$ influences learning, guiding us to improve it.

2.2 Analysing Information Gain and improving self-training

Various works argue that SSL cannot gain information about the label generating process $p(y|x)$ solely from unlabeled data without priors or assumptions [5,20,4]. Then, *how* does this framework benefit learning? A rigorous theory does not exist. We conjecture that information useful for learning is given to the system via a *subtle assumption* when pseudo-labels $Y_T^\star$ are created from class-posteriors. We analyse this hypothesis through the lens of Information Gain (IG). IG is defined as the change of entropy H of a system from a prior to a new state due to a new condition: $IG = H_{old} - H_{new}$. Predicted posteriors for sample x_i have entropy $H_i = -\sum_{c=1}^{C} p(c|x_i;\theta') \log p(c|x_i;\theta')$. A 1-hot posterior has 0 entropy. When a pseudo-label is created via $y_{T,i}^\star = \arg\max_c p(c|x_i;\theta')$, an *assumption* is made that the class with highest posterior is correct and its posterior *should* be 1. This generates $IG_i = H_i - 0 = H_i$ bits of info. If the prediction is correct, the pseudo-label matches the true unknown label, $y_{T,i}^\star = y_{T,i}$, and the obtained H_i bits of information will facilitate learning. If the prediction is wrong, $y_{T,i}^\star \neq y_{T,i}$, the system is given H_i bits of wrong information that hinder learning. We show next how confidence calibration relates to IG and how gains can be maximized.

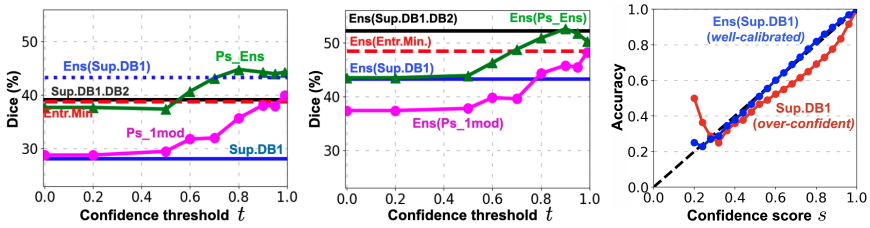
A model's confidence score $s \in [0, 1]$ is considered *well calibrated* if $s \approx \frac{\sum_i \mathbb{1}[y_i'=y_i]\mathbb{1}[s_i'=s]}{\sum_i \mathbb{1}[s_i'=s]} \forall s$ (for discretized s). That is, confidence s is equal to the ratio of correct ($y_i' = y_i$) over all predictions made with confidence $s = s_i'$. As confidence score we use the model's maximum class posterior, $s_i = \max_c p(y=c|x_i;\theta)$ [11]. In this case, if we *assume* model confidence is well-calibrated, we can estimate the *expected* IG given to the system *on average* when a sample predicted with confidence s is assigned a 1-hot pseudo-label: $IG_{i,s} = sH_i - (1-s)H_i$. We plot this function in Fig. 1a. Standard neural nets, commonly used with self-training, are over-confident in practice [9,16,10]. Therefore we also plot the function $IG_{i,s} = \delta s H_i - (1-\delta s)H_i$, where δ models the level of over or under-confidence. Lower δ represents lower percentage of correct predictions than estimated by s (over-confidence). Fig. 1a shows that expected IG_i is positive for samples above a certain confidence and negative below. This supports the common practice of using confidence threshold t when creating pseudo-labels. Importantly, we find that for over-confident models, not only this threshold increases, but also the expected positive IG_i decreases! This means that in self-training, using well-calibrated or under-confident models a) leads to higher gains and better learning than with over-confident models; b) allows a wider range of values for threshold t without injecting negative IG to the system. The latter is practically important as choosing hyper-parameters is difficult for methods using unlabeled data [19].

Guided by the above findings, we improve the framework by introducing the currently most reliable method for obtaining calibrated uncertainty with neural networks, model ensembling [16,13]. Fig. 1b shows the framework with network ensembles. Individual models are trained with different weight initialization and data sampling, found sufficient for better uncertainty estimates (Fig. 2c). Ensembling is also known to improve predictive accuracy, benefiting quality of pseudo-labels. We will demonstrate in Sec. 3 empirical evidence aligned with our above two theoretical observations: Self-training via ensembling a) provides performance gains for different values of t the magnitude of which follows trend aligned with what our theoretical analysis suggests (Fig. 1a), b) provides high performance for a wide range of t making configuration easier and more reliable.

3 Experimental evaluation

3.1 Data and model configuration

We evaluate transduction on multi-class TBI segmentation in the following setup.
DB1: This database was acquired in one clinic using 3 scanners. Consists of 180 subjects who underwent MRI including T1w, T2w, FLAIR, and SWI or GRE (SWI & GE used interchangeably here). It includes $\sim$85% moderate/severe and $\sim$15% mild cases. DB1 labels are used for training in all following experiments.
DB2: This includes 101 subjects, acquired in 9 clinics with 4 scanner models, all different from DB1. T1w, T2w, FLAIR and SWI scans are available for all patients. DB2 consists of $\sim$35% moderate/severe and $\sim$65% mild cases. These distribution shifts between DB1 and DB2 make for a challenging benchmark. DB2 is used as *target* data D_T in following experiments, except for those in Sec. 3.4, where its labels are used for training and evaluation is done on DB3.
DB3: This database includes scans of 265 subjects from 14 sites (9 overlap with DB2, 5 new). Sequences and severity are similar to DB2. This database was unlabeled, kept unseen during model development. We used DB3 for blinded comparison of induction versus transduction after model development (Sec. 3.4).
Manual annotations: Experts annotated any abnormality visible in DB1 and DB2. Here, we consider 6 classes: core, oedema, petechial hemorrhages, intraventricular hemorrhages, non-TBI lesions, monitoring probe.
Pre-processing: We registered each image to the corresponding T1w, resampled to 1mm isotropic resolution, and performed z-score intensity normalization.
Main model: We use a 3D Convolutional Network (CNN), DeepMedic, previously used for TBI segmentation [14,18]. We use the 'wide' variant with default hyper-parameters (from: `https://github.com/deepmedic/deepmedic`, v0.8.4).
Compared methods and configuration: Besides self-training, we explore transduction with Entropy Minimization (EM), originally developed for inductive SSL [8]. Moreover, training a CNN on pseudo-labels by ensemble with no confidence threshold t (Ps_Ens_0 below) is equal to ensemble-distillation on unlabelled data [3]. Striving for reliability, we chose the studied methods for their



(a) Single models (b) Ensembles (c) Calibration

Fig. 2: (a,b) Average Dice (%) over all 6 TBI classes achieved by different methods. Shown as function of confidence threshold t for self-training methods. (c) Reliability diagram [9] of a single CNN *Sup.DB1* and an ensemble *Ens(Sup.DB1)*.

simplicity, to avoid hyper-parameter tuning that is impractical in SSL [19]. Except t that we study below, the only hyper-parameter is the weight of unsupervised vs supervised cost, which we set to 1 for self-training and EM.

3.2 Comparing supervised and transductive learning

Each experiment below is repeated for 10 seeds. Fig. 2 summarizes results. Shown is mean Dice over all classes and seeds, as space is limited for per-class analysis.
Superv.DB1: The supervised baseline are 10 models trained with full supervision on DB1, evaluated on DB2. Models achieve 28.1% Dice on average over all classes. The low score is due to the tiny size of mild TBI lesions where few false positives greatly impact Dice (Fig. 3b), and domain shift between DB1-DB2.
Superv.DB1.DB2: We also evaluate how well a supervised model would perform if labels for DB2 were available for training. We train 10 models using 100% DB1 and 80% DB2 labels, and evaluate on remaining 20% of DB2. We repeat for 5 folds of DB2. SSL without labels on DB2 is not expected to outperform this, but hopefully approach it.
Entr.Min.: We perform transductive learning by training 10 models with supervision over DB1 and entropy minimization [8] over unlabelled DB2 images. We evaluate predictions made on DB2 at end of training. Transduction with EM improves over baseline *Superv.DB1* (Fig. 2a).
Ps_1m_t: We evaluate transduction via self-training using a single CNN as f_θ. We use confidence threshold t on posteriors of a supervised CNN (*Superv.DB1*) to create pseudo-labels, and re-train a CNN on the extended training set. We repeat for 10 models of *Superv.DB1* and 7 values of t to investigate its effect. Using $t=0$ equals to using segmentations of *Superv.DB1* as pseudo-labels because no prediction was less confident. Fig. 2 shows improvements over *Superv.DB1*. Performance reaches EM only for best choice of t. High t values are required, aligned with our theoretical findings about over-confident CNNs (Fig. 1a).
Pseudo-labels from ensemble: We assess performance of a CNN when trained with pseudo-labels derived from confidence estimates of an ensemble. We make

ensemble **Ens(Superv.DB1)** combining 10 baseline CNNs (*Superv.DB1*). Fig. 2c shows the ensemble is indeed well-calibrated. We obtain pseudo-labels using threshold t on ensemble's posteriors for DB2. We re-train 10 CNNs using DB1 labels and DB2 pseudo-labels. Transductive predictions are obtained. We show average Dice of 10 seeds as **Ps_Ens**_t. We repeat for varying t. CNN retrained using ensemble segmentations as pseudo-labels (*Ps_Ens*_0, $t = 0$, ala ensemble distillation [3]) outperform *Ps_1m*_t, thanks to the better segmentation of ensemble, which is not surprising. Most importantly, treating confidence via threshold t offers performance gains that follow trend similar to what we derived for IG with well-calibrated models (Fig. 1a), supporting our theoretical results. Close-to-optimal performance is achieved with wide range of t, in contrast to *Ps_1m*_t, explained by the wider range of t with positive IG of the former according to our findings (Fig. 1a). This facilitates the choice of a reliable t value. Finally, without labels for DB2, *Ps_1mod*_t outperforms inductive *Sup.DB1.DB2*, likely thanks to direct optimization of predictions on DB2 by transductive learning.

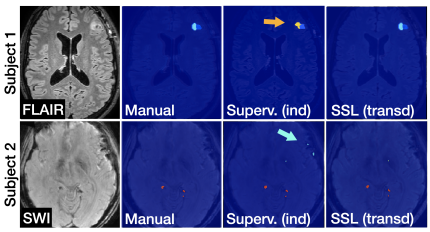
Results by ensembles: If ensemble is used to make pseudo-labels, then effective self-training should improve an ensemble. To assess this, we average transductive predictions from 10 models *Ps_Ens*_t per t, obtaining transductive predictions of self-trained ensemble **Ens(Ps_Ens**_t**)** (Fig. 1b). We compare with ensembles of CNNs re-trained with pseudo-labels from single CNNs (*Ens(Ps_1mod*_t*)*), entropy minimization (*Ens(Entr.Min)*) and supervision on DB1∪DB2 (*Ens(Superv.DB1.DB2)*). Fig. 2b shows that self-training improves the ensemble, with its transductive results (*Ens(Ps_Ens*_t*)*) exhibit lower and less stable performance. Importantly, this holds for all values $t > 0.5$, aligned with our findings for well-calibrated models (Fig. 1a). In contrast, ensembles re-trained on pseudo-labels from individual CNNs (*Ens(Ps_1mod*_t*)*) reach the baseline ensemble only for optimal t. Transductive ensemble trained via entropy minimization also surpasses the inductive ensemble. The best transductive ensemble approaches inductive *Ens(Superv.DB1.DB2)* that requires labels for DB2.

3.3 Comparing inductive and transductive semi-supervised learning

We compare transduction and induction with the same model after self-training. To do this, we divide DB2 in 2 equal folds. We start with pseudo-labels for DB2 from *Ens(Sup.DB1)* (Sec. 3.2). For 10 seeds, we perform self-training using 100% DB1 labels and 50% pseudo-labeled DB2. After training, we obtain transductive predictions for the DB2 fold used in self-training and inductive predictions for the other fold. We switch DB2 folds and repeat, to obtain these predictions for whole DB2 (*Ps_ens_t*). Finally, we ensemble inductive and transductive predictions (separately) from 10 seeds, to assess ensemble performance (*Ens(Ps_Ens*_t*)*). We repeat for $t = 0.7$ and 0.9 to assess along the wider effective range. Results are shown in Table 3a. Transductive SSL outperforms inductive SSL in all cases.

Method	Ind	Transd
Sup.DB1	28.1	-
Ps_Ens_0.7	39.7	43.3
Ps_Ens_0.9	41.9	44.8
Ens(Sup.DB1)	43.3	-
Ens(Ps_Ens_0.7)	48.8	50.8
Ens(Ps_Ens_0.9)	50.0	53.3

(a) Inductive vs transductive SSL (Dice%)



(b) Example predictions

Tab.3a) Comparison of transduction and induction. We show average Dice% over all TBI classes and 10 seeds. Differences (Ind/Transd) are significant ($p<0.05$). Fig.3b) Example predictions by induction (*Ens(Sup.DB1)*) and transduction (*Ens(Ps_Ens_0.7)*). Yellow arrow shows edema mistaken as related to monitoring probe. Cyan arrow shows false-positive TBI core. Both corrected by transductive SSL. Mild TBI lesions are often tiny, hence Dice degrades even with small errors.

Table 1: Dice% between predicted and manually refined lesions on DB3. Transduction overlaps more than induction ($p<0.05$ for average over classes, *Avg*).

Method	Core	Oed.	Non-TBI	Probe	Petech.	Intrav.	Avg.
Induction, fully supervised:	75.7	77.9	84.4	85.7	53.4	83.9	76.8
Transduction, self-training:	79.7	79.8	91.9	82.4	58.8	83.9	79.4

3.4 Blinded comparison via manual refinement of segmentations

SSL is often found beneficial when labels are limited but its reliability is still questioned in practical settings, where supervised learning is preferred, e.g. large studies with more labels (Sec. 1). To address this, we compare transductive SSL with supervised training when using all 281 labeled cases of DB1∪DB2, largest segmented cohort reported for MRI TBI, and predict 265 cases of multi-center DB3, which were unlabeled and held-out during model development for objective evaluation. We perform blinded comparison via manual segmentation refinement.

Because manual refinement is time consuming, we defined a process that requires only one refinement per image. We obtain *inductive* predictions on DB3 using *Ens(Superv.DB1.DB2)*, which is supervised on $DB1 \cup DB2$ (Sec. 3.2). From these predictions we make pseudo-labels with $t = 0.7$ (choice based on Fig. 1a). Using $DB1 \cup DB2$ labels and $DB3$ pseudo-labels we train an ensemble, similar to *Ens(Ps_Ens_0.7)* (Sec. 3.2), and obtain *transductive* predictions. We then create an 'in-between' segmentation per image by averaging inductive and transductive posteriors (ala ensembling). These latter segmentations were manually refined by clinicians with TBI expertise to meet the highest standards for follow-up studies on TBI phenotyping. Experts were unaware how the predictions were created.

Finally, we evaluated overlap of manually-refined segmentations with inductive and transductive predictions. Table 1 shows the results. Transduction clearly

out-performs inductive supervised learning. These results show that transductive SSL can facilitate large-scale studies by reducing effort for manual refinement.

4 Conclusion

We explored the potential of transduction for segmentation. We showed with extensive experiments that if test data are available in advance, transduction outperforms induction by supervision or SSL. We also presented theoretical and empirical evidence that using well-calibrated or under-confident models facilitates self-training. Future research should evaluate transduction on other data and perform analysis with multiple metrics, beyond Dice, to capture the method's effect more comprehensively. Having set the first stone with this study, future works should explore transductive segmentation with other SSL methods. We believe these results will inspire further research in transductive SSL, which is well-suitable for retrospective medical image analysis studies.

5 Acknowledgements

This work received funding from the UKRI London Medical Imaging & Artificial Intelligence Centre for Value Based Healthcare. VFJN is funded by an Academy of Medical Sciences / The Health Foundation Clinical Scientist Fellowship. DKM is supported by the National Institute for Health Research (NIHR, UK) through the Cambridge NIHR Biomedical Research Centre, and by a European Union Framework Program 7 grant (CENTER-TBI; Grant agreement 602150).

References

1. Bai, W., Oktay, O., Sinclair, M., Suzuki, H., Rajchl, M., Tarroni, G., Glocker, B., King, A., Matthews, P.M., Rueckert, D.: Semi-supervised learning for network-based cardiac mr image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 253–260. Springer (2017)
2. Boudiaf, M., Masud, Z.I., Rony, J., Dolz, J., Piantanida, P., Ayed, I.B.: Transductive information maximization for few-shot learning. arXiv preprint arXiv:2008.11297 (2020)
3. Bucilua, C., Caruana, R., Niculescu-Mizil, A.: Model compression. In: Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining. pp. 535–541 (2006)
4. Castro, D.C., Walker, I., Glocker, B.: Causality matters in medical imaging. Nature Communications 11(1), 1–10 (2020)
5. Chapelle, O., Scholkopf, B., Zien, A.: Semi-supervised Learning. MIT Press, Cambridge, Mass., USA (2006)
6. Cui, W., Liu, Y., Li, Y., Guo, M., Li, Y., Li, X., Wang, T., Zeng, X., Ye, C.: Semi-supervised brain lesion segmentation with an adapted mean teacher model. In: International Conference on Information Processing in Medical Imaging. pp. 554–565. Springer (2019)

7. De, J., Li, H., Cheng, L.: Tracing retinal vessel trees by transductive inference. BMC bioinformatics 15(1), 1–20 (2014)
8. Grandvalet, Y., Bengio, Y., et al.: Semi-supervised learning by entropy minimization. In: CAP. pp. 281–296 (2005)
9. Guo, C., Pleiss, G., Sun, Y., Weinberger, K.Q.: On calibration of modern neural networks. In: International Conference on Machine Learning. pp. 1321–1330. PMLR (2017)
10. Hein, M., Andriushchenko, M., Bitterwolf, J.: Why relu networks yield high-confidence predictions far away from the training data and how to mitigate the problem. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 41–50 (2019)
11. Hendrycks, D., Gimpel, K.: A baseline for detecting misclassified and out-of-distribution examples in neural networks. arXiv preprint arXiv:1610.02136 (2016)
12. Huang, R., Noble, J.A., Namburete, A.I.: Omni-supervised learning: scaling up to large unlabelled medical datasets. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 572–580. Springer (2018)
13. Jungo, A., Balsiger, F., Reyes, M.: Analyzing the quality and challenges of uncertainty estimations for brain tumor segmentation. Frontiers in neuroscience 14, 282 (2020)
14. Kamnitsas, K., Ledig, C., Newcombe, V.F., Simpson, J.P., Kane, A.D., Menon, D.K., Rueckert, D., Glocker, B.: Efficient multi-scale 3d cnn with fully connected crf for accurate brain lesion segmentation. Medical image analysis 36, 61–78 (2017)
15. Kervadec, H., Dolz, J., Granger, É., Ayed, I.B.: Curriculum semi-supervised segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 568–576. Springer (2019)
16. Lakshminarayanan, B., Pritzel, A., Blundell, C.: Simple and scalable predictive uncertainty estimation using deep ensembles. Advances in Neural Information Processing Systems (2017)
17. Lee, D.H., et al.: Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In: Workshop on challenges in representation learning, ICML. vol. 3 (2013)
18. Monteiro, M., Newcombe, V.F., Mathieu, F., Adatia, K., Kamnitsas, K., Ferrante, E., Das, T., Whitehouse, D., Rueckert, D., Menon, D.K., et al.: Multiclass semantic segmentation and quantification of traumatic brain injury lesions on head ct using deep learning: an algorithm development and multicentre validation study. The Lancet Digital Health 2(6), e314–e322 (2020)
19. Oliver, A., Odena, A., Raffel, C.A., Cubuk, E.D., Goodfellow, I.: Realistic evaluation of deep semi-supervised learning algorithms. In: Advances in Neural Information Processing Systems. pp. 3235–3246 (2018)
20. Schölkopf, B., Luo, Z., Vovk, V.: Empirical inference: Festschrift in honor of vladimir n. vapnik (2013)
21. Scudder, H.: Probability of error of some adaptive pattern-recognition machines. IEEE Transactions on Information Theory 11(3), 363–371 (1965)
22. Vapnik, V.: Statistical Learning Theory. A Wiley-Interscience publication, Wiley (1998), `https://books.google.co.uk/books?id=GowoAQAAMAAJ`
23. Wang, Z., Zhu, X., Adeli, E., Zhu, Y., Zu, C., Nie, F., Shen, D., Wu, G.: Progressive graph-based transductive learning for multi-modal classification of brain disorder disease. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 291–299. Springer (2016)

24. Yu, L., Wang, S., Li, X., Fu, C.W., Heng, P.A.: Uncertainty-aware self-ensembling model for semi-supervised 3d left atrium segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 605–613. Springer (2019)
25. Zheng, H., Zhang, Y., Yang, L., Liang, P., Zhao, Z., Wang, C., Chen, D.Z.: A new ensemble learning framework for 3d biomedical image segmentation. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 33, pp. 5909–5916 (2019)