Consistency-based Semi-supervised Evidential Active Learning Framework for Robust Classification of Radiology Images

Shafa Balaram, Manh Cuong Nguyen, Yang Yu, Ivan Ho Mien, Huijuan Yang, Navya Sindhu Vanjavaka, Nurul Hafidzah Rahim, Alistair Giap Chuan Koh, Shaheen Low, Ashish Jith Sreejith Kumar, Feri Guretno, Yih Yian Sitoh, Ashraf Kassim*, *Senior Member, IEEE*, and Pavitra Krishnaswamy*

***Abstract*— Deep learning offers high performance for radiology image classification, but relies on large, expert-annotated datasets. Semi-supervised learning and active learning approaches can leverage unlabelled samples and mitigate the annotation burden. Combining these techniques via semi-supervised active learning (SSAL) can compound their benefits, yet the effectiveness of this approach may be hindered by unreliable uncertainty estimation and consistency enforcement. To address these challenges, we propose Consistency-based Semi-supervised Evidential Active Learning (CSEAL). CSEAL is a principled SSAL framework leveraging evidential learning for reliable estimation of predictive uncertainty for consistency enforcement and prioritised sampling. First, we develop evidential counterparts of leading semi-supervised methods with different consistency enforcement mechanisms: Pseudo-labelling, Virtual Adversarial Training, Mean Teacher, and NoTeacher, and demonstrate customisability of CSEAL. Second, we introduce Noise Robust-evidential NoTeacher, an enhancement over evidential NoTeacher, that uses consensus principles and a small-loss inclusion mechanism to learn with noisy annotations. In extensive experiments on many X-ray, CT, and MRI datasets, we demonstrate that CSEAL offers substantial performance gains over competitive SSAL baselines and enhances robustness in noisy annotation scenarios. Third, we translate CSEAL into an annotation platform and demonstrate its value as an aid for real-world radiology image annotation. Specifically, our CSEAL-assisted platform performs close to fully-supervised learning with very small labelled datasets, enables accurate auto-labelling, and maintains performance even when only noisy labels from junior annotators are available for training. Our work offers new opportunities to enhance the efficiency of clinical image annotation and model development workflows.**

Research funding and infrastructure from the A*STAR Singapore International Graduate Award scholarship, A*STAR Institute for Infocomm Research, and the AME Programmatic Funds (A20H6b0151).

Shafa Balaram is with the National University of Singapore, Singapore 117583. Shafa Balaram, Manh Cuong Nguyen, Yang Yu, Ivan Ho Mien, Huijuan Yang, Alistair Giap Chuan Koh, Shaheen Low, Ashish Jith Sreejith Kumar, Feri Guretno, Pavitra Krishnaswamy are with the Institute for Infocomm Research, Agency for Science, Technology and Research (A*STAR), Singapore 138632. Ivan Ho Mien, Navya Sindhu Vanjavaka, Nurul Hafidzah Rahim, Yih Yian Sitoh are with the National Neuroscience Institute, Singapore 308433. Ashraf Kassim is with the Singapore University of Technology and Design, Singapore 487372.

*Equal contribution.

✉ shafa.balaram@u.nus.edu, pavitrak@a-star.edu.sg

***Index Terms*— Active learning, Medical image annotation, Noisy labels, Semi-supervised learning, Theory of Evidence**

I. Introduction

DEEP learning models offer cutting-edge performance in radiology image classification tasks [1]. However, training such models requires large-scale datasets which have been labelled by clinical experts. In practice, annotating medical images is a resource-intensive process and the curation of extensive, labelled datasets is often not viable. New approaches are emerging to reduce the annotation burden. For instance, a targeted labelled set can be constructed using the most informative samples [2]. Alternatively, the considerable number of unlabelled samples acquired routinely in clinical databases can also be used for model development [3].

Active learning (AL) provides a human-in-the-loop paradigm to incrementally query the most informative images using an acquisition function [2]. The selected samples are labelled by an *oracle* or human expert and used for model development. This process is iterated until a target performance or annotation budget is reached. However, models trained on limited sets of labelled samples may overfit, leading to the *cold-start problem*, where initial biases are propagated in subsequent rounds [4].

Semi-supervised learning (SSL) leverages large unlabelled datasets alongside small labelled sets during training to improve model performance [3]. SSL methods can learn the underlying data structure from unlabelled samples and reduce the labelling requirements [5]. However, these methods typically construct the labelled sets via random sampling and do not consider the feature distributions, class distributions, and informativeness of the labelled samples used for training.

Since AL and SSL are complementary approaches, there have been efforts to combine them to compound their benefits and mitigate their drawbacks [4], [6], [7]. Many such semi-supervised active learning (SSAL) approaches are *consistency-based*, meaning that predictions on unlabelled samples are matched to their targets, or informativeness of unlabelled samples is determined based on the consistency of predictions across augmented versions. In these SSAL approaches, the

semi-supervised component relies on point estimates of probabilities (e.g., softmax) to encourage consistency, but errors can propagate when the full distributions of the prediction probabilities are not accounted for. Further, the active learning component can be undermined by mismatch between the softmax probabilities used to estimate informativeness and the true predictive uncertainty [8]. These technical challenges are more pronounced at lower labelling budgets, where limited supervision constrains model training. Additionally, the utility of SSAL approaches in addressing challenges of real-world radiology image annotation settings is yet to be demonstrated.

A particular challenge for real-world radiology image annotation tasks is that they often entail labelling errors [9]. In this domain, the accuracy of human interpretation and the resulting labels is impacted by image acquisition protocols, artifacts, as well as other subjective factors – resulting in inter-rater variability. Although iterative annotation based on multi-rater consensus can improve label quality, it is highly resource-intensive and often infeasible. In turn, noisy annotations can undermine typical assumptions in AL and SSL methods, and hamper model performance and generalisability [10], [11]. While prior AL and SSL works have individually addressed noisy annotations [12], [13], there is limited work on noisy annotations within the SSAL literature.

Motivated by the above challenges, we develop Consistency-based Semi-supervised Evidential Active Learning (CSEAL). CSEAL is a principled SSAL framework to learn and leverage the predictive uncertainty alongside the classification objective in an end-to-end manner. Our approach leverages evidential learning [14] to map image inputs to a distribution over the prediction probabilities, uses this distribution to enforce prediction consistency via a novel semi-supervised loss formulation, and simultaneously estimates predictive uncertainty for active learning. Beyond addressing challenges of reliable uncertainty estimation in SSAL, our modified CSEAL loss can also be used to address noisy annotations. The CSEAL methodology is applicable across image modalities. We focus experiments on three key radiology modalities. We interleave CSEAL within a practically relevant assisted annotation process and demonstrate its value as an aid for real-world radiology image annotation tasks. Our main contributions are as follows:

- We introduce CSEAL, a unified evidential framework for SSAL. CSEAL is formulated to support customisability to different SSL and AL approaches. Notably, we derive the evidential counterparts of four leading consistency enforcement mechanisms: Pseudo-labelling, Virtual Adversarial Training, Mean Teacher, and NoTeacher. CSEAL also enables uncertainty estimation to prioritise informative samples for labelling.
- We perform extensive experiments with 2D and 3D radiology image datasets spanning X-ray, CT and MRI modalities [15]–[17]. We demonstrate that CSEAL achieves over 90% of the fully supervised AUROC with less than 10% labelling budget. Further, CSEAL offers substantial performance gains over three state-of-the-art SSAL methods, especially at lower labelling budgets and for classes with lower prevalence.
- To enhance robustness to label noise, we propose the semi-supervised Noise Robust-evidential NoTeacher (NR-eNoT) method. Specifically, we train two evidential networks using a joint loss derived from the consensus principle of multi-view learning. In addition to unlabelled samples, NR-eNoT identifies and incorporates small-loss labelled samples for parameter updates. We show that NR-eNoT enables effective learning in noisy annotation scenarios [18], and improves performance over the naive evidential NoTeacher method by up to 3.7% in AUPRC.
- We develop the CSEAL-assisted Medical Expert Annotation Platform (CSEAL-MEAP) and demonstrate it on a real-world clinical brain image annotation task. We show that our platform halves the labelling budget required to achieve 90% of the fully-supervised performance and enables more accurate auto-labelling than conventional active learning. Further, even when trained solely with noisy annotations from junior annotators, our approach achieves a performance comparable to training with clean annotations from senior annotators.

This work builds upon and substantially extends a preliminary version of CSEAL from our previous conference paper [19] with several novel contributions. First, we revamp the loss function with a cross-entropy-based prediction term aligned with the probabilistic nature of classification. We also perform an independent scaling of the regularisation term for enhanced learning of evidences, which provides flexibility in emphasising the prediction accuracy or uncertainty modelling. The new loss enables generalisability across all three key radiology imaging modalities (X-ray, CT, and MRI). Second, we showcase framework adaptations for both 2D and 3D input data, and expand the evaluation to 5 datasets and 3 baselines. Third, we introduce a versatile noise-robust SSAL approach, and demonstrate it under varying degrees of label noise and in practical annotation scenarios. Finally, we translate CSEAL into an annotation platform and demonstrate its performance in a real-world annotation task to showcase value-add in reducing many barriers for in-house annotation. Through a theoretically-grounded, customisable, and practical framework, our work offers enhanced capabilities for large-scale clinical image annotation and label-efficient model development in radiology.

II. Related Work

A. Active Learning

Active learning (AL) aims to select informative samples for labelling based on certain criteria such as uncertainty, diversity, or hybrid combinations of both. Classical uncertainty-based approaches [20] make use of model posterior probabilities, and may not fully represent the data distribution. To address this, diversity methods such as core-set [21]; and hybrid approaches such as Batch Active learning by Diverse Gradient Embeddings (BADGE) [22] and combined uncertainty-diversity methods [23] have been proposed. However, these methods require distance computations or clustering steps which can be computationally expensive for large datasets. Other uncertainty quantification techniques, such as Monte Carlo dropout [24] and evidential deep learning [14], [25],

have also been applied to medical images. We note that using these AL techniques within supervised learning paradigms can be prone to overfitting when labelled data is sparse (e.g., initial annotation rounds). This pertains to the cold-start problem, where biases from the early annotation rounds are propagated to the subsequent rounds [4].

B. Semi-Supervised Learning

Semi-supervised learning (SSL) addresses the annotation burden by leveraging unlabelled data [3]. Pseudo-labelling [26] trains a supervised model with labelled samples, uses the supervised model to predict 'pseudo-labels' for unlabelled samples, and then utilises both the labelled and pseudo-labelled sets during training. This approach is vulnerable to the propagation of label noise [27]. Virtual Adversarial Training [28] enhances model robustness to adversarial perturbations by enforcing isotropic smoothness in the target label distribution using a local distributional smoothness loss. Mean Teacher [29] is a consistency-based method using a student-teacher pair to enforce consistent predictions across augmentations. NoTeacher [30] was proposed to handle the confirmation bias due to EMA in Mean Teacher [31]. While these SSL methods have been successfully applied to medical images, they construct the labelled sets via random sampling, without considering the feature distributions, class distributions, and informativeness of the labelled samples used for training.

C. Semi-Supervised Active Learning

Semi-supervised active learning (SSAL) has been studied mainly in computer vision. Cost-Effective Active Learning (CEAL) combined high-confidence pseudo-labelling with uncertainty-based strategies [32], but is prone to confirmation bias. MixMatch was integrated with AL methods such as label propagation [33] and data summarisation [34], but its reliance on MixUp augmentation hinders medical imaging applications. Recent approaches focus on consistency and augmentation-awareness. For instance, Temporal and Cyclic Output Discrepancy (TOD+COD) compares prediction variation across training steps [6], but suffers from prediction error propagation, also known as confirmation bias. VAT+AugVar quantifies informativeness via prediction variance on augmented samples [4], but is highly dependent on augmentation and multiple forward passes. SSAL has also been explored in medical imaging, UniSAL integrates pseudo-labeling, contrastive learning, and uncertainty-based sample selection, but its rigid design limits generalization, especially on highly imbalanced medical datasets [35]. Pseudo-labelling with BADGE demonstrated effectiveness across biomedical datasets [36] but uses a computationally expensive k-means clustering step. BoostMIS combines adaptive pseudo-labelling, augmentation-aware consistency, and label propagation for spinal MRI [7], selecting samples based on instability to perturbations and weighted entropy. However, BoostMIS requires setting of multiple hyperparameters and depends on class-balanced sampling. Overall, despite advances in consistency-based SSAL approaches [37]–[39], the use of point estimates to encourage consistency could limit effective sample prioritisation for labelling and propagate errors during training.

D. Noise-Robust Learning

There are advances in the AL, SSL and SSAL literature to improve learning in the presence of noisy labels. In AL, [12] addressed noisy pseudo-labels in histopathology imaging. However, this single-network approach is vulnerable to a self-confirmation loop. Further, annotation noise has been addressed by crowd-sourcing labels [40], but obtaining labels from numerous annotators with specialised domain expertise is impractical in clinical applications. In SSL, robust training techniques have been employed to distinguish between noisy labels for unsupervised learning and clean labels for supervised learning [13]. However, these methods do not leverage large unlabelled datasets and their success would rely on the accuracy of pseudo-labels. Further, [41] combined Pseudo-labelling with confidence estimation and contrastive learning to address annotation noise, but did not perform evaluations under real-world noisy annotation scenarios. For SSAL, there are methods to handle noisy pseudo-labels. [42] introduced an example re-weighting AL method that used a pool attention module and an unsupervised, symmetric soft cross-entropy loss within a Mean Teacher-based framework, but these approaches require a clean expert-labelled set (i.e., no annotation noise). To our knowledge, there are no studies on SSAL with only noisy annotations.

E. Translation into Annotation Workflows

The medical imaging community has advanced software platforms to reduce the annotation burden for efficient model development [43]. These platforms typically incorporate conventional AL and offer automated labelling functionalities to assist annotators [44]. However, they do not utilise unlabelled data to reduce the annotation burden during model training. To our knowledge, the efficacy of SSAL approaches is yet to be demonstrated in real-world annotation workflows.

III. Methods

A. Problem formulation

An input sample x can represent a 2D image (or slice), or a series of images in a 3D scan, and is associated with a target y for l classes (or abnormalities), each assigned a binary label $k \in \{-, +\}$. When $l = 1$, y is uni-label; when $l > 1$ it is multi-label. In each annotation round, the training set contains L_{Tr} labelled samples $\{\mathbf{x}_i^{\mathrm{L}}, \mathbf{y}_i\}_{i=1}^{L_{Tr}}$ and U unlabelled samples, $\{\mathbf{x}_i^{\mathrm{U}}\}_{i=1}^{U}$. We also consider the scenario where the target may not align with ground-truth annotations $\mathbf{y}$ and denote these noisy annotations as $\hat{\mathbf{y}}$. In Subsection III-B, we introduce CSEAL: a principled and unified framework to derive the evidential counterpart of consistency-based semi-supervised learning methods and combine it with uncertainty-based active learning. In Subsection III-C, we provide an extension of CSEAL to robustly address noisy annotation scenarios. Fig. 1 illustrates the CSEAL framework for binary classification ($l = 1$). For multi-label classification ($l > 1$), each network uses multiple binary classification heads.

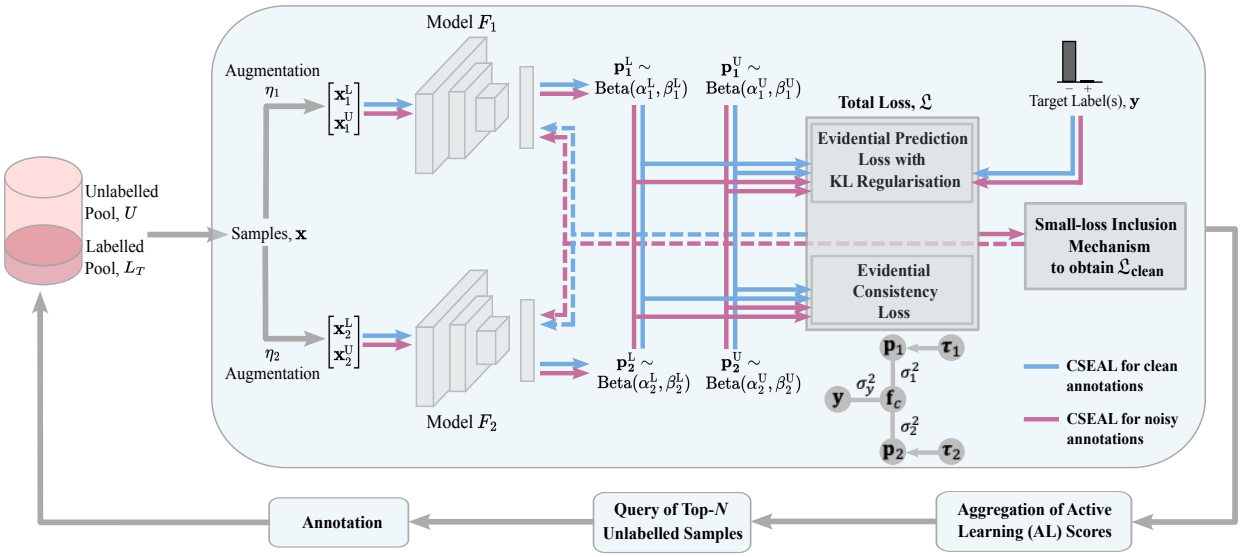

Fig. 1. Overview of the Consistency-based Semi-supervised Evidential Active Learning (CSEAL) framework with clean annotations for evidential NoTeacher and noisy annotations for Noise Robust-evidential NoTeacher. Input samples $\mathbf{x}$ (images or scans) are augmented by $\{\boldsymbol{\eta_1}, \boldsymbol{\eta_2}\}$. The transformed samples $\{\mathbf{x_1}, \mathbf{x_2}\}$ are fed to the evidential networks $\{\boldsymbol{F_1}, \boldsymbol{F_2}\}$, and mapped to Beta distribution prior parameters of their label predictors $\{\mathbf{p_1}, \mathbf{p_2}\}$. The overall CSEAL loss combines prediction, regularisation, and consistency losses, and includes a small-loss inclusion mechanism to exclude noisy annotations from backpropagation. Solid and dashed arrows indicate the forward pass and backpropagation steps respectively. After training, the inferred evidential active learning scores are used to prioritise $\boldsymbol{N}$ unlabelled samples for annotation before the process is repeated. The framework is customisable for different network structures (single/dual), consistency targets, and active learning strategies (see evidential Pseudo-labelling, evidential Virtual Adversarial Training, evidential Mean Teacher).

B. Overview of the CSEAL framework

Motivated by the need to reliably estimate prediction probabilities and their associated uncertainties for semi-supervised active learning, we leverage an evidential learning strategy. Evidential deep learning [14] applies the Dempster-Shafer Theory of Evidence [45] and the theory of Subjective Logic [46] to quantify the predictive uncertainties by modelling binary predictions as subjective opinions characterised by Beta distributions [14]. Beta distributions have the conjugate prior property enabling closed form expressions for updating belief masses with new evidence. We integrate evidential learning into semi-supervised deep learning to introduce the Consistency-based Semi-supervised Evidential Active Learning (CSEAL) framework. CSEAL consists of two stages: 1) training an evidential semi-supervised learning classifier, and 2) querying unlabelled samples using evidential active learning to expand the labelled set. The two stages form one loop, and the loops are iteratively repeated until the annotation budget is exhausted.

1) Semi-supervised learning component: Given a batch of labelled and unlabelled samples, $\mathbf{x} = \{\mathbf{x}^{\text{L}}, \mathbf{x}^{\text{U}}\}$, transformation functions η_1 and η_2 are applied to generate augmented samples $\mathbf{x}_1 = \{\mathbf{x}_1^{\text{L}}, \mathbf{x}_1^{\text{U}}\}$ and $\mathbf{x}_2 = \{\mathbf{x}_2^{\text{L}}, \mathbf{x}_2^{\text{U}}\}$. These are fed into convolutional networks F_1 and F_2 to produce logits $\mathbf{f_1}$ and $\mathbf{f_2}$. In standard binary classification, the logits are mapped to label predictors $\mathbf{p}_1 = [p_1^-, p_1^+]^\top$ and $\mathbf{p}_2 = [p_2^-, p_2^+]^\top$ using a sigmoid function. The Bernoulli variables $\mathbf{p}_1$ and $\mathbf{p}_2$ are point estimates of the first-order probabilities. This point estimation may be mismatched with the true predictive uncertainty, and can contribute to error propagation when enforcing consistency. To address this challenge, CSEAL estimates a density over probability assignments and thus effectively models the second-order probabilities.

Specifically, CSEAL employs evidential deep learning principles [14], [25] to model priors over $\mathbf{p}_1$ and $\mathbf{p}_2$ as Beta distributions parameterised by $\boldsymbol{\tau}_1 = [\alpha_1, \beta_1]$ and $\boldsymbol{\tau}_2 = [\alpha_2, \beta_2]$. These parameters are computed as $\boldsymbol{\tau} = \text{Softplus}(\mathbf{f}) + \mathbf{1}$, mapping logits to label-wise evidences using the Softplus activation, ensuring α and β are greater than 1. The term *evidence* refers to the support collected for positive or negative predictions. CSEAL uses an evidential loss for semi-supervised model training, comprising of prediction loss, regularisation loss, and consistency loss components, derived below.

The prediction loss $\mathfrak{L}_{\text{pred}}$ is the Bayes risk of the binary cross-entropy loss between $\mathbf{y}$ and $\mathbf{p}_1$:

$$\mathfrak{L}_{\text{pred}}(\mathbf{p}_1, \mathbf{y}) = \int_0^1 \text{CE}(p_1^k, y_k)\ \text{Beta}(\alpha_1, \beta_1)\ d\mathbf{p}_1\ . \quad (1)$$

The regularisation loss term is included to penalise high-uncertainty predictions through a Kullback-Leibler (KL) divergence term. For labelled samples, the Beta distribution priors are adjusted as $\dot{\boldsymbol{\tau}}_1 = \mathbf{y} + (1 - \mathbf{y}) \odot \boldsymbol{\tau}_1$, and the KL term quantifies the divergence between a uniform Beta distribution and the adjusted prior. This penalises evidences for non-target labels. The resulting regularisation loss is:

$$\mathfrak{L}_{reg}(\mathbf{p}_1, \mathbf{y}) = \text{KL}\left[\text{Beta}(\mathbf{p}_1|\dot{\boldsymbol{\tau}}_1)\ \|\ \text{Beta}(\mathbf{p}_1|\mathbf{1})\right]. \quad (2)$$

The consistency loss $\mathfrak{L}_{\text{cons}}$ of two label predictors $\mathbf{p}_1, \mathbf{p}_2$ is defined as the Bayes risk of the squared error between them:

$$\mathfrak{L}_{\text{cons}}(\mathbf{p}_1, \mathbf{p}_2) = \iint_0^1 \|\mathbf{p}_1 - \mathbf{p}_2\|^2\ \text{Beta}(\alpha_1, \beta_1)\ \text{Beta}(\alpha_2, \beta_2)\ d\mathbf{p}_1 d\mathbf{p}_2, \quad (3)$$

where the prediction probabilities $\hat{\mathbf{p}}_1$ and $\hat{\mathbf{p}}_2$, also used during inference, are calculated as the Beta distribution mean. That is, $\hat{\mathbf{p}} = [\hat{p}^+, \hat{p}^-]^\top = [\alpha/E, \beta/E]^\top$, with total evidence $E = \alpha + \beta$.

Overall, the CSEAL loss for the l^{th} class comprises of the above three terms:

$$\mathfrak{L}_{\text{CSEAL}}(\mathbf{x}_1, \mathbf{y}) = \lambda_{sup}\ \mathfrak{L}_{\text{pred}}(\mathbf{p}_1, \mathbf{y}) + \lambda_t\ \mathfrak{L}_{\text{reg}}(\mathbf{p}_1, \mathbf{y}) + \lambda_{cons}\mathfrak{L}_{\text{cons}}(\mathbf{p}_1, \mathbf{p}_2), \tag{4}$$

where λ_t is an adaptive regularisation coefficient to independently scale the regularisation term for enhanced evidence learning. It also supports balancing the prediction accuracy with uncertainty modelling. λ_{sup} and λ_{cons} are scaling constants to enable customisation to different consistency enforcing mechanisms, and $\mathbf{p}_1$ is a function of $\mathbf{x}_1$ as noted above.

From Equation (4), the supervised baseline with evidential learning (eSUP) optimises the following loss on labelled data:

$$\mathfrak{L}_{\text{eSUP}}(\mathbf{x}_1^{\text{L}}, \mathbf{y}) = \lambda_{sup}\ \mathfrak{L}_{\text{pred}}(\mathbf{p}_1^{\text{L}}, \mathbf{y}) + \lambda_t\ \mathfrak{L}_{\text{reg}}(\mathbf{p}_1^{\text{L}}, \mathbf{y}), \tag{5}$$

where λ_t is ramped up linearly from 0 over the first T_r epochs. We now instantiate the CSEAL framework on four SSL methods: namely Pseudo-labelling [26], Virtual Adversarial Training [28], Mean Teacher [29], and NoTeacher [30]. The evidential counterparts of these methods are described below.

Evidential Pseudo-labelling (ePSU): This approach has two stages. First, an eSUP model is trained with labelled samples to infer pseudo-labels, $\mathbf{y}_{psu}$, for the unlabelled samples $\mathbf{x}_1^{\text{U}}$ based on the mean of the Beta distribution prior. Second, pseudo-labels are recomputed at each weight update to further guide the eSUP loss. Alongside the eSUP loss, a binary cross-entropy loss ensures the updated pseudo-labels align with the eSUP model predictions:

$$\mathfrak{L}_{\text{ePSU}}(\mathbf{x}_1, \mathbf{y}) = \mathfrak{L}_{\text{eSUP}}(\mathbf{x}_1^{\text{L}}, \mathbf{y}) + \mathfrak{L}_{\text{pred}}(\mathbf{p}_1^{\text{U}}, \mathbf{y}_{psu}). \tag{6}$$

Evidential Virtual Adversarial Training (eVAT): To increase robustness of the evidential networks to small perturbations or noise, we define a local distributional smoothness loss between the Beta distribution priors of the original features of the unlabelled samples $\mathbf{x}_1^{\text{U}}$ and those perturbed by uniform random noise. Specifically, this loss is computed as the Bayes risk of the squared error between the label predictor $\mathbf{p}_1^{\text{U}}$ and its adversarial counterpart $\mathbf{p}_{1,adv}^{\text{U}}$:

$$\mathfrak{L}_{\text{eVAT}}(\mathbf{x}_1, \mathbf{y}) = \mathfrak{L}_{\text{eSUP}}(\mathbf{x}_1^{\text{L}}, \mathbf{y}) + \lambda_{cons}\ \mathfrak{L}_{\text{cons}}(\mathbf{p}_1^{\text{U}}, \mathbf{p}_{1,adv}^{\text{U}}). \tag{7}$$

Evidential Mean Teacher (eMT): Augmented samples $\mathbf{x}_1$ and $\mathbf{x}_2$ are input into student and teacher evidential networks. The student model's label predictors are used to compute the evidential supervised loss for labelled samples. The unsupervised consistency loss is the Bayes risk of the squared error between the label predictors of the student and teacher models, $\mathbf{p}_1$ and $\mathbf{p}_2$, based on their Beta distribution priors. Both losses are backpropagated through the student model.

$$\mathfrak{L}_{\text{eMT}}(\mathbf{x}, \mathbf{y}) = \mathfrak{L}_{\text{eSUP}}(\mathbf{x}_1^{\text{L}}, \mathbf{y}) + \lambda_{cons}\ \mathfrak{L}_{\text{cons}}(\mathbf{p}_1, \mathbf{p}_2). \tag{8}$$

Evidential NoTeacher (eNoT): eNoT uses a probabilistic graphical model (PGM), shown in Fig. 1, where $\mathbf{p}_1$, $\mathbf{p}_2$, and $\mathbf{y}$ are conditionally independent random variables. Observed variables $\boldsymbol{\tau}_1$, $\boldsymbol{\tau}_2$, and $\mathbf{y}$ are distinct nodes, while $\mathbf{p}_1$ and $\mathbf{p}_2$ are linked to a consensus function $\mathbf{f}_c \in [\mathbf{0}, \mathbf{1}]$. $\mathbf{f}_c$ enforces agreement between $\mathbf{p}_1$ and $\mathbf{p}_2$ for all samples and maps them to $\mathbf{y}$ for labelled data. Following [30], the differences between $\mathbf{y}$, $\mathbf{p}_1$, $\mathbf{p}_2$, and $\mathbf{f}_c$ are modelled as zero-centered Gaussian distributions parameterised by σ_y, σ_1, and σ_2. For an input batch, we express the NoTeacher loss likelihood in terms of the label predictors, then compute the Bayes risk as follows:

$$\mathfrak{L}_{\text{eNoT}}(\mathbf{x}, \mathbf{y}) = \lambda_{sup,1}\ \mathfrak{L}_{\text{eSUP}}(\mathbf{x}_1^{\text{L}}, \mathbf{y}) + \lambda_{sup,2}\ \mathfrak{L}_{\text{eSUP}}(\mathbf{x}_2^{\text{L}}, \mathbf{y}) + \lambda_{cons}^{\text{L}}\ \mathfrak{L}_{\text{cons}}(\mathbf{p}_1^{\text{L}}, \mathbf{p}_2^{\text{L}}) + \lambda_{cons}^{\text{U}}\ \frac{n_U}{n_L}\mathfrak{L}_{\text{cons}}(\mathbf{p}_1^{\text{U}}, \mathbf{p}_2^{\text{U}}). \tag{9}$$

For ePSU and eVAT, we use a single-network architecture with F_1 only, omitting F_2. For eMT and NoTeacher eNoT, we employ two separate networks. When $\mathbf{x}$ is a volumetric scan, the classification backbone is adapted for depth information using a modified 2D-CNN (see Subsection IV-F.1).

2) Active learning component: Prioritising the most uncertain samples to augment the labelled training set can enhance model performance. To this end, we leverage the underlying evidential mathematical framework of CSEAL that learns the predictive uncertainty (PU) from both labelled and unlabelled samples in Stage 1. Implicitly, PU captures both the aleatoric (data) and epistemic (model) uncertainties, and is estimated through a single forward pass during inference. We define PU for an unlabelled sample x^{U} using the label-wise evidences and therefore, the Beta distribution parameters, as follows:

$$\text{PU}(\text{x}^{\text{U}}) = 1 - \frac{\alpha^{\text{U}} - 1}{\alpha^{\text{U}} + \beta^{\text{U}}} - \frac{\beta^{\text{U}} - 1}{\alpha^{\text{U}} + \beta^{\text{U}}}. \tag{10}$$

For multi-label scenarios, label-wise uncertainties are averaged to compute an image-level score, and images with the highest scores are selected for annotation to augment the training set. Similar to [25], we use dropout for efficient parameter sampling, robust uncertainty estimation, and training stability. When $\mathbf{x}$ represents slices from volumetric scans, training a 2D classifier is less resource-intensive than a 3D one. However, active learning must prioritise scan-level samples due to slice contiguity, reducing redundancy between labelled and unlabelled pools. Subsection IV-F.1 details the mapping of slice-level to scan-level scores.

C. NR-eNoT: Noise-robust extension of eNoT

To enable robust learning with noisy annotations during training, we extend CSEAL's evidential NoTeacher method. Our approach is inspired by the use of dual-network structures to reduce error accumulation [47], [48]. Specifically, we leverage the different learning abilities and agreement between networks F_1 and F_2 on augmented views of the same samples to identify noisy annotations. We posit that small loss values correspond to "clean" labels [47] and incorporate a *small-loss inclusion mechanism* that uses the sum of prediction and consistency losses from both networks to distinguish samples with clean vs. noisy annotations. We term this method Noise Robust-Evidential NoTeacher (NR-eNoT).

NR-eNoT computes the loss derived from the PGMs shown in Fig. 1 twice: 1) to filter noisy annotations in the labelled set

$\mathbf{x}^{\mathrm{L}}$ and exclude them from backpropagation, and 2) to optimise model weights using only samples with clean labels.

In the first step, the noisy loss $\mathfrak{L}_{\mathrm{noisy}}$ is calculated for all the labelled samples in a mini-batch:

$$\mathfrak{L}_{\mathrm{noisy}}(\mathbf{x},\hat{\mathbf{y}}) = \lambda_{sup,1}\ \mathfrak{L}_{\mathrm{pred}}(\mathbf{p}_1^{\mathrm{L}},\hat{\mathbf{y}}) + \lambda_{sup,2}\ \mathfrak{L}_{\mathrm{pred}}(\mathbf{p}_2^{\mathrm{L}},\hat{\mathbf{y}}) + \lambda_{cons}^{\mathrm{L}}\ \mathfrak{L}_{\mathrm{cons}}(\mathbf{p}_1^{\mathrm{L}},\mathbf{p}_2^{\mathrm{L}}). \quad (11)$$

The first two terms of (11) are inspired by the implementation of Co-teaching [47], except for the fact that we have combined supervised losses from both networks in a single backpropagation update. The additional term is the consistency term $\mathfrak{L}_{\mathrm{cons}}$, inspired by multi-view learning principles [49], that drives the two networks to align their predictions. A smaller $\mathfrak{L}_{\mathrm{cons}}$ indicates higher agreement between the networks. Hence, a smaller noisy loss indicates higher likelihood of correct labels. To retain samples with the smallest loss values i.e., cleaner annotations, we define a remembrance rate R_t [47]. To avoid overfitting to noisy data in later epochs, we gradually decrease R_t from 1 to $1-\gamma$ (γ is the drop rate).

Algorithm 1 Training of the NR-eNoT model

Input: Labelled training set $\{\mathbf{x}_i^{\mathrm{L}},\hat{\mathbf{y}}_i\}_{i=1}^{L_T}$, unlabelled training set $\{\mathbf{x}_i^{\mathrm{U}}\}_{i=1}^{L_U}$, networks F_1 and F_2 with learning rate ξ, drop rate γ, epoch T_k, total epochs T_{total}, number of iterations N
Initialise $R_{t=1}$ to 1
for $t=1$ to T_{total} **do**
  for $n=1$ to N **do**
    Fetch a mini-batch $\{\mathbf{x}_i^{\mathrm{L}},\hat{\mathbf{y}}_i,\mathbf{x}_i^{\mathrm{U}}\}_{i=1}^{|B|}$
    Apply transformations η_1 and η_2 to $\mathbf{x}_1$ and $\mathbf{x}_2$
    $\mathbf{f}_1 = F_1(\mathbf{x}_1|\theta_1)$ and $\boldsymbol{\tau}_1 = \mathrm{Softplus}(\mathbf{f}_1)+\mathbf{1}$
    $\mathbf{f}_2 = F_2(\mathbf{x}_2|\theta_2)$ and $\boldsymbol{\tau}_2 = \mathrm{Softplus}(\mathbf{f}_2)+\mathbf{1}$
    Calculate $\mathfrak{L}_{\mathrm{noisy}}$ (Eq. (11)) using $\boldsymbol{\tau}_1^{\mathrm{L}}$ and $\boldsymbol{\tau}_2^{\mathrm{L}}$
    Rank samples based on $\mathfrak{L}_{\mathrm{noisy}}$ and retain R_t of labelled samples with the smallest values
    Calculate $\mathfrak{L}_{\mathrm{clean}}$ (Eq. (12))
    Update $\theta_1 = \theta_1 - \xi\nabla\mathfrak{L}_{\mathrm{clean}}$
    Update $\theta_2 = \theta_2 - \xi\nabla\mathfrak{L}_{\mathrm{clean}}$
  end for
  Update $R_t = 1-\min\left\{\frac{t}{T_k}\gamma,\gamma\right\}$
end for

In the second step, retained clean samples and unlabelled data are used to compute the following semi-supervised loss:

$$\mathfrak{L}_{\mathrm{clean}}(\tilde{\mathbf{x}}^{\mathrm{L}},\mathbf{x}^{\mathrm{U}},\tilde{\mathbf{y}}) = \lambda_{sup,1}\ \mathfrak{L}_{\mathrm{eSUP}}(\tilde{\mathbf{x}}_1^{\mathrm{L}},\tilde{\mathbf{y}}) + \lambda_{sup,2}\ \mathfrak{L}_{\mathrm{eSUP}}(\tilde{\mathbf{x}}_2^{\mathrm{L}},\tilde{\mathbf{y}}) + \lambda_{cons}^{\mathrm{L}}\ \mathfrak{L}_{\mathrm{cons}}(\tilde{\mathbf{p}}_1^{\mathrm{L}},\tilde{\mathbf{p}}_2^{\mathrm{L}}) + \lambda_{cons}^{\mathrm{U}}\ \frac{n_U}{R_t}\mathfrak{L}_{\mathrm{cons}}(\mathbf{p}_1^{\mathrm{U}},\mathbf{p}_2^{\mathrm{U}}), \quad (12)$$

where $\sim$ denotes samples estimated to be correctly labelled, with $\tilde{\mathbf{x}}$ being a subset (size R_t) of inputs deemed correctly labelled by the two evidential networks. Equation (12) incorporates the KL regularisation loss term $\mathfrak{L}_{\mathrm{reg}}$ to enhance evidence learning for this subset. The training process of NR-eNoT is outlined in Algorithm 1. In noisy annotation scenarios, determining informativeness may require active learning approaches beyond predictive uncertainty (see Subsection IV-B). Table I summarises the key characteristics, advantages, typical application scenarios and active learning methods studied for the different CSEAL methods.

IV. Experimental Setup

We describe three sets of experiments: 1) to evaluate CSEAL performance against semi-supervised active learning baselines, 2) to investigate the robustness of CSEAL in scenarios with noisy annotations, and 3) to demonstrate the relevance of CSEAL in a real-world annotation task.

A. Datasets

We benchmarked the performance of CSEAL on four publicly available datasets, which were chosen to represent scenarios spanning 2D and 3D scans, diverse radiology imaging modalities, and uni- and multi-labelled classification tasks.

1) NIH-14 Chest X-Ray Dataset [15]: This dataset contains 112,120 frontal radiographs labelled for 14 abnormalities using natural language processing (NLP) on radiological reports. Of images in this dataset, 46.1% are abnormal and 40.1% are multi-labelled. We utilised the publicly available training (70%), validation (10%) and test (20%) splits which do not have patient overlaps [50].

2) RSNA Brain CT Dataset [16]: We started with the RSNA 2019 Intracranial Haemorrhage Detection Challenge (Stage 1) dataset and retained a single study per patient. This led to a dataset of 17,079 scans and 588,732 slices. Each slice is annotated by a neuroradiologist to indicate the presence or absence of 5 sub-types of intracranial haemorrhage. There are 14.2% abnormal slices, out of which 30.1% are multi-labelled. We randomly split the dataset into 60% training, 20% validation and 20% testing without patient overlaps. We pre-processed the images as per the leading solution of the RSNA Challenge [51]. This includes converting raw pixel values to Hounsfield Units, (b) windowing around the width and centre values read from the DICOM files, (c) normalising pixel intensities to [0, 255], and (d) resizing slices to 256×256. In addition to the 2D classification task, we also address the 3D task, where a scan is considered positive for a given class if any of its slices contains the abnormality.

TABLE I
OVERVIEW OF CSEAL METHODS. ALL METHODS CATER TO LIMITED LABEL SCENARIOS. 'RESOURCES' REFERS TO COMPUTATIONAL RESOURCES.

	ePSU	eVAT	eMT	eNoT	NR-eNoT
Characteristics	Refines pseudo-labels	Adversarial training	Dual-network	Dual-network	Loss-based filtering
Advantages	Easy implementation	Perturbation smoothness	Smoothes fluctuations	↓ confirmation bias	↑ Noise robustness
Applications	Low resources	Perturbed input scenarios	Low resources	High resources	Annotation errors
AL investigated	PU	PU	PU	PU	PU/AU/BADGE

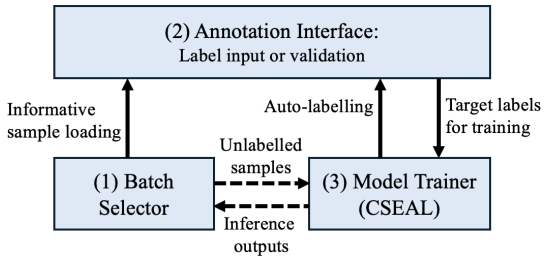

Fig. 2. Schematic of the CSEAL-assisted Medical Expert Annotation Platform (CSEAL-MEAP) illustrating the flow of information between its various modules.

3) NoisyCXR Dataset [18]: This benchmark dataset of 26,684 chest radiographs contains pneumonia-like opacities with noisy annotations simulated via a realistic process. Following the implementation of [18], we randomly flipped set proportions of NLP-derived labels for *unclear cases* (positive for the "Consolidation" or "Infiltration" but not "Pneumonia") to introduce different noise levels, and assigned clean expert labels for the remaining images. For our study, we set the randomly flipped proportion to [0%, 30%, 50%, 70%] to simulate annotators with overall noise rates of [10.2%, 17.2%, 31.4%, 40.7%], respectively. As per [18], we allocated 50% of the images for training (mix of clean and noisy annotations) and 50% for evaluation (only clean expert labels).

4) Knee MRNet Dataset [17]: This dataset contains 1370 knee MRI exams, each from 3 physical views. All scans were manually labelled, based on reports, for Meniscal and Anterior Cruciate Ligament (ACL) tears. 80.6% of these 3D scans contain at least one category of tear and we considered the binary classification task of any abnormality (Abnormal) in the axial plane. Similar to [11], we utilised the official training (1130 scans) and test (120 scans) splits without patient overlaps. We further generated a validation pool with 20% of the training set through stratified sampling following [17].

In line with [17], no additional pre-processing steps are applied for the 3D classification tasks in the RSNA Brain CT [16] and Knee MRNet datasets [17].

B. Baselines

First, we compared CSEAL to three state-of-the-art (SOTA) semi-supervised active learning methods: TOD+COD [6], VAT+AugVar [4], and BoostMIS [7]. For 3D scan classification, to the best of our knowledge, there are no existing 3D baselines for comparisons. Therefore, we adapted the three SOTA SSAL methods in a manner consistent with our CSEAL variants. We also included a fully-supervised baseline trained using the maximum labelling budget of 100% and confirmed that its accuracy is comparable (within 1.5%) to that of typical fully-supervised benchmarks [11], [30], [51]. We denote the CSEAL approaches and baselines by their semi-supervised and active learning methods, e.g., eNoT+PU combines evidential NoTeacher with evidential predictive uncertainty.

Second, we characterised the performance of our proposed NR-eNoT approach in relation to that of eNoT+PU (semi-supervised active learning without noise-robust learning), eNoT+Random (semi-supervised learning only), eSUP+PU (supervised active learning only [14]), and the Joint Training with Co-Regularization (JoCoR) baseline [48] (supervised noise-robust learning only). Notably, JoCoR applies the consensus principle from multi-view learning [49] by training two networks with a joint loss, and shares a similar objective with eNoT, though it is fully supervised. Further, we examined the influence of different active learning strategies on CSEAL performance in noisy annotation scenarios. Beyond the predictive uncertainty strategy (Equation (10)), we also performed experiments using evidential aleatoric uncertainty (AU) [19] to capture the data uncertainty, and Batch Active Learning by Diverse Gradient Embeddings (BADGE) [22] to query more diverse sets of highly uncertain samples.

C. Realistic Active Sampling Process

For all experiments, we designed an active sampling process that reflects practical clinical annotation workflows. First, our active sampling process neither stratifies nor balances the class distributions, so it is more realistic for class-imbalanced scenarios. Second, we allocated only a small portion of the labelling budget for validation compared to training, i.e., $L_V \ll L_T$. We initialised L_{Tr} and L_V by random sampling until the annotation budget is met while ensuring complete class coverage for each label. We further grow the validation set progressively using randomly selected samples to maintain the $L_{Tr} : L_V$ ratio at different labelling budgets. This process precludes an ideal alignment of the training and validation class distributions. We also focus on the low labelling budget regime of 0.25%–35% across datasets. Overall, our experimental design constitutes a realistic but challenging setup.

D. CSEAL-assisted Medical Expert Annotation Platform

To incorporate CSEAL in guiding clinical annotators effectively using semi-supervised active learning, we developed the CSEAL-assisted Medical Expert Annotation Platform (CSEAL-MEAP). As shown in Fig. 2, CSEAL-MEAP provides a human-in-the-loop paradigm and comprises three modules: (1) a *batch selector* to prioritise informative studies using active learning and load the next batch for annotation; (2) an *annotation interface* for the annotator to input, validate, or correct labels; and (3) a *model trainer* that trains a semi-supervised learning model in parallel as more labels become available to enhance the estimation of informativeness and the accuracy of predicted labels for auto-labelling.

E. Real-world Annotation Task

We evaluated CSEAL-MEAP on the task of annotating a real-world dwMRI dataset [52] containing 296 diffusion-weighted brain studies (8304 slices) acquired in the axial plane. The relevant institutional review board (SingHealth CIRB Ref 2019/2804) approved extraction and use of the data with a waiver of informed consent. Abnormal regions with DWI hyperintensities for each slice were manually delineated by neuroradiology fellows (senior annotators) and radiography students (junior annotators). Thus, the junior and senior annotators provided a set of noisy and clean (ground-truth) labels

for each image, respectively. From the seniors' labels, 7.36% of the slices were abnormal. The juniors disagreed with their seniors in 3.95% of all slices.

We examined how CSEAL-MEAP could enhance annotation of dwMRI dataset through a proof-of-concept study conducted under four practically relevant annotation workflows. Workflow A(SUP+Random) represents a human-only paradigm with randomly-loaded studies and supervised model training. Workflow B(SUP+EU) follows conventional active learning with epistemic (model) uncertainty estimated using Monte-Carlo dropout [24]. Hence, A and B correspond to typical workflows in annotation platforms [43], [44]. Both C(eNoT+PU) and D(NR-eNoT+PU) employ CSEAL methods, with D designed to address noisy annotations. All workflows involve loading batches with pre-obtained labels, training models, generating results, and auto-labelling within the platform. We evaluated whether CSEAL-MEAP can (1) reduce annotation burden (workflows A–C), (2) increase auto-labelling accuracy (A–C), and (3) maintain performance despite noisy annotations on real-world clinical data (workflows C–D).

F. Implementation Details

We now provide details on model training and evaluation.

1) Classification Backbones: For all methods, we chose the classification backbones of prior works, where available: DenseNet121 for NIH-14 Chest X-ray [53], DenseNet169 for RSNA Brain CT (2D) [51], an enhanced version of MRNet with AlexNet for both RSNA Brain CT (3D) and Knee MRNet [30], ResNet50 for NoisyCXR [18] and ResNet18 for dwMRI. We tested the ResNet50 and Swin Transformer backbones in RSNA Brain CT (2D), and chose the backbone with the best performance. We initialised all classifiers with pre-trained ImageNet weights, except for NoisyCXR where we randomly initialised weights [18]. After each annotation round, the model weights are reset for re-training. For evidential methods, we applied a dropout rate of 0.5 before the last fully-connected layer to estimate evidences. For each class, we then mapped the logits of the negative and positive labels to the respective Beta distribution parameters $\boldsymbol{\tau} = [\alpha, \beta]$ using a Softplus activation function. For volumetric scans, CSEAL uses a parameterised 2D CNN that condenses the features extracted from slices through indexed max-pooling [30] for scalability and computational efficiency. For slice-level classification tasks with 3D datasets, we prioritise scans for active learning to reflect typical clinical annotation practice. We take the 95$^{\text{th}}$ percentile of uncertainty scores across slices to obtain the scan-level scores as it reduces skewness to outlier values.

2) Training Losses for Baselines: For all baselines, we used multi-label binary cross-entropy as the supervised loss term [30]. Where applicable, we used the same consistency loss function as the original method. For VAT+AugVar and the Adversarial Uncertainty Selector stage of BoostMIS, we computed the multi-label Kullback-Leibler (KL) divergence for the LDS loss, as employed by [30] for the evaluation of VAT for multi-label classification.

3) Optimisation: For all datasets and methods, we used an Adam optimiser ($\beta_{\text{Adam}} = [0.9, 0.999]$, $\epsilon_{\text{Adam}} = 1 \times 10^{-8}$) with a linear decay scheduler that scales the learning rate by 0.1 based on the validation loss. We saved the best checkpoint based on the class-averaged validation metric. To ensure fair comparison of semi-supervised against supervised methods, we used equal batch sizes of labelled and unlabelled samples such that each training epoch covers a complete round of the labelled samples. We also maintained an EMA copy for other single-network approaches in experiments involving TOD+COD and Mean Teacher variants. All approaches were implemented in PyTorch 1.4.0 or higher. Since a complete grid search for optimal hyperparameters is computationally infeasible, we adopted configurations from [4], [6], [7], [30] for the baselines, and tuned key hyperparameters (details provided in Table S1). We also describe the augmentations used in Supplement Section B.

4) Computational Complexity: We analysed CSEAL's computational complexity by considering the computational overhead associated with its two main stages of SSL and AL. Given the number of AL training iterations κ, the number of epochs ζ, the number of classes l, the complexities of forward and backward passes for a backbone with M parameters denoted by $\mathcal{F}_M$ and $\mathcal{B}_M$, the computational complexity of the CSEAL approaches is $O\left(\kappa n_U \left(l + \mathcal{F}_M\right) + \kappa\zeta n \left(\mathcal{F}_M + l + \mathcal{B}_M\right)\right)$. Here, n_U is the number of unlabelled samples, and $n = n_L + n_U$ is the total sample count. We provide the derivation in Supplement Subsection A. Thus, the computational costs depend on the model re-training frequency, dataset type (2D or 3D, image dimensions, and n), and model architecture (M). To accelerate the runtime, we trained the models on NVIDIA GeForce RTX 2080 Ti GPUs (11GB memory) with parallel processing. The total time for CSEAL training and inference ranges 12–16 hours across different methods for the RSNA Brain CT (2D) dataset and is 3 hours for NR-eNoT for the NoisyCXR dataset, generally comparable to those for several competing methods (TOD+COD, BoostMIS, JoCoR, DivideMix).

G. Model Inference

For performance metrics, we report the best test AUROC across models for dual-network methods and its corresponding AUPRC (given the class imbalance). For multi-label classification tasks, we compute metrics for each class and report the average. We repeat each experiment with three seeds for initialising random parameters, and report mean and standard deviation across seeds. For approaches with two identical networks, such as evidential NoTeacher, we compute scores using the model with the best validation AUROC. We consider a method to have higher annotation efficiency if it achieves better performance with fewer labelled samples.

V. Results and Analysis

A. The CSEAL Framework with Clean Annotations

1) CSEAL outperforms the SOTA baselines: Fig. 3 (a-b, d-e, g-h) and Fig. S1 show the average test AUROC and AUPRC as a function of labelling budget for the NIH-14 Chest X-ray, RSNA Brain CT (2D, 3D) and Knee MRNet datasets. Overall, CSEAL outperforms the SSAL baselines with ePSU+PU,

eMT+PU and eNoT+PU emerging consistently among the top-performing methods across datasets. At the mid-range labelling budgets (denoted by arrows), our best performing method, namely eNoT+PU, offers gains of 2.4%–19.7% in AUROC and 1.2%–17.5% in AUPRC across baselines and datasets. We observe that the augmentation-aware and adversarial-based approaches, BoostMIS, VAT+AugVar and eVAT+PU, have larger standard deviations for RSNA Brain CT (3D) and Knee MRNet. Regardless, the general trends hold and the CSEAL methods have higher annotation efficiencies. Notably, eNoT+PU reaches 90% and 70% of the fully supervised AUROC and AUPRC performance respectively with less than 10% of the labelling budget, i.e., fewer than 4360 chest radiographs, 136 and 273 brain CT scans in 2D and 3D, respectively, and 125 knee MR scans. We posit that CSEAL's gains across datasets are likely due to its evidential learning-based formulation that models second-order probabilities for uncertainty estimation, rather than point-estimates.

2) CSEAL improves performance of low prevalence classes: Fig. 3(c, f, i) provide the AUROC gains for each method and class over the lowest performing baseline at the mid-range labelling budgets. Classes are ordered by increasing prevalence in the test set, e.g., from Hernia (least frequent) to Infiltration (most frequent) in the NIH-14 Chest X-ray. For 2D tasks, CSEAL methods substantially outperform all three SOTA baselines, especially for low prevalence classes: eNoT+PU improves over TOD+COD by 0.6%–9.2% for classes with prevalence $< 3.5\%$ (Hernia to Pleural Thickening), and by 4.6%–17.3% for classes with prevalence $< 5\%$ (Epidural to Subarachnoid). CSEAL remains competitive for more prevalent classes (e.g., Atelectasis, Subdural). In 3D, eMT+PU and eNoT+PU show substantial gains of up to 8.2%–16.5% across classes and baselines. These gains may be due to their dual-network structure and consistency enforcement mechanisms.

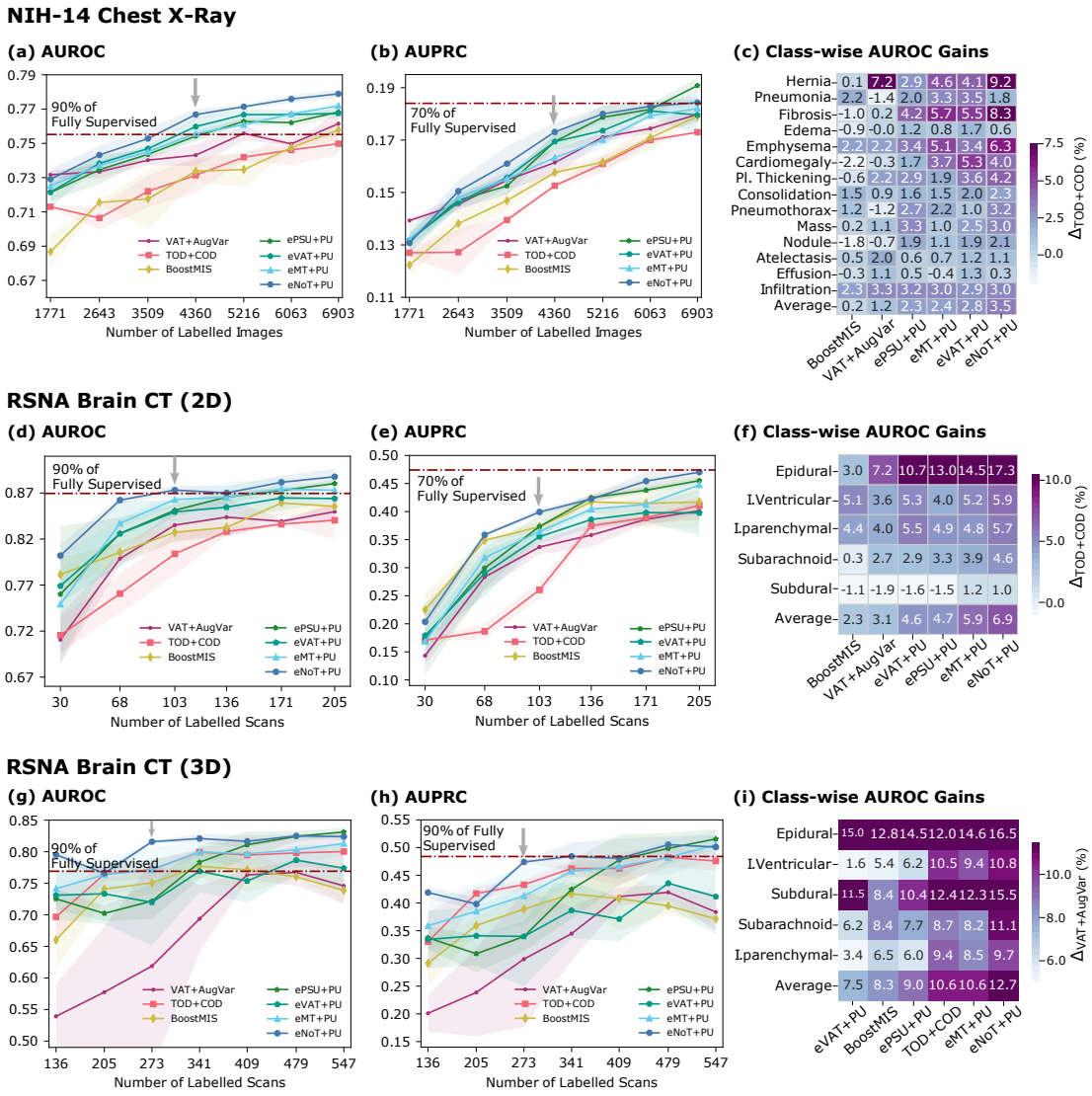

Fig. 3. Performance results of CSEAL framework on the NIH-14 Chest X-ray and RSNA Brain CT datasets. Our evidential Pseudo-labelling (ePSU), Virtual Adversarial Training (eVAT), Mean Teacher (eMT) and NoTeacher (eNoT) methods combined with predictive uncertainty (PU) are compared against three semi-supervised active learning baselines, VAT+AugVar, TOD+COD and BoostMIS. (a, d, g) Test AUROC and (b, e, h) Test AUPRC at different labelling budgets. Plots show mean $\pm$ standard deviation. (c, f, i) Class-wise breakdown of AUROC gains over the baseline with the lowest performance at the mid-range labelling budgets denoted by the arrows. Methods on x-axes are ordered by their average performance.

3) Ablation studies: Our CSEAL framework combines semi-supervised learning (SSL), active learning (AL) and evidential learning (EDL). We ablated these components to assess their impact on performance in the RSNA Brain CT (2D) dataset. Table II reports the average test performance of the best performing method, eNoT+PU, at selected labelling budgets.

a) Leveraging unlabelled samples with SSL: Learning the underlying information from unlabelled samples contributes to eNoT+PU outperforming eSUP+PU [14] by 1.6%–3.7% in AUROC and 5.2%–6.2% in AUPRC across budgets, especially with fewer labelled samples. The substantial gains in AUPRC suggest that low prevalence classes benefit more from the utilisation of unlabelled samples during training.

b) Prioritised sampling with AL: eNoT+PU gains 1.2%–3.8% in AUROC and 11.3%–11.7% in AUPRC over eNoT+Random across labelling budgets. Hence, including slices with higher predictive uncertainty in the labelled set improves classification performance of true positives. This observation aligns with the expectation that predictive uncertainty is higher for abnormal slices, as they have lower prevalence than healthy slices. Further, the effectiveness of our uncertainty-based sampling in low labelling regimes suggests CSEAL's potential to mitigate the cold-start problem.

c) Parameterisation using EDL: Modelling the predictive uncertainty during training does not appear to directly improve the performance when comparing eNoT+Random against NoT [30]. However, the objective of using evidential networks is to enable explicit uncertainty estimation for AL. This, in turn, enables a performance boost of eNoT+PU over NoT.

TABLE II
ABLATION STUDIES FOR ENOT+PU IN THE RSNA BRAIN CT (2D) DATASET. THE PARTS ABLATED ARE SEMI-SUPERVISED LEARNING (ESUP+PU), ACTIVE LEARNING (ENOT+RANDOM) AND EVIDENTIAL LEARNING (NOT). THE FIRST AND SECOND ROWS REPORT THE AVERAGE TEST AUROC AND AUPRC RESPECTIVELY AT DIFFERENT LABELLING BUDGETS.

Method	Number of Scans (Labelling budget in %)		
	68 (0.50)	136 (1.00)	205 (1.50)
eSUP+PU	82.54 ± 1.96 30.65 ± 2.37	85.16 ± 0.46 36.30 ± 0.54	87.16 ± 1.03 40.87 ± 0.62
eNoT+Random	82.39 ± 1.20 24.18 ± 2.40	85.50 ± 1.25 30.78 ± 1.05	87.62 ± 0.84 35.78 ± 3.52
NoT	82.07 ± 1.33 26.08 ± 1.22	$\mathbf{87.64 \pm 0.83}$ 35.93 ± 1.37	86.57 ± 1.60 37.30 ± 2.15
eNoT+PU	$\mathbf{86.23 \pm 0.39}$ $\mathbf{35.86 \pm 0.35}$	87.01 ± 0.72 $\mathbf{42.24 \pm 0.61}$	$\mathbf{88.77 \pm 0.89}$ $\mathbf{47.03 \pm 1.06}$

B. The CSEAL Framework with Noisy Annotations

1) NR-eNoT demonstrates higher robustness to label noise: Fig. 4 compares the performance of Noise-Robust evidential NoTeacher (NR-eNoT) combined with various sampling strategies against four baselines (eSUP+PU, JoCoR, eNoT+Random, eNoT+PU) under four different noise rates. For the noise rates of 10.2%, 17.2% and 31.4%, NR-eNoT generally outperforms all baselines, with gains of up to 7.6% in AUPRC across different noise rates and labelling budgets. Although JoCoR improves over eSUP+PU in some settings, our evidential approaches across noise rates and labelling budgets improve over JoCoR and eSUP+PU in general. Further, NR-eNoT tends to outperform eNoT with up to 6.7% higher AUPRC – this highlights the importance of accounting for noisy annotations. Finally, for noise rates upto 31.4%, NR-eNoT outperforms JoCoR by up to 4.2% AUPRC, demonstrating the impact of leveraging unlabelled samples for noise robustness. For the extremely high 40.7% noise rate, NR-eNoT is at best comparable to JoCoR (under random sampling), suggesting that the gains from including unlabelled samples diminish at extreme noise levels. However, we found that the prevalence of ambiguous to obvious lesions in the set of samples misclassified by the model generally follows that of the original test set.

Among active sampling strategies, NR-eNoT performs best with BADGE, offering AUPRC gains of up to 3.7% over other active sampling methods at the lower noise rates of 10.2% and 17.2%. This is potentially due to its ability to trade-off between sampling uncertain and diverse images. BADGE is followed closely by AU, which selects up to 3.1% more ambiguous samples (i.e., samples where some annotators delineated opacities but the consensus was "No Pneumonia-like opacity" [18]) than other AL strategies. AU has a lower computational complexity than BADGE and presents a scalable alternative for larger datasets. However, at the higher noise rates of 31.4% and 40.7%, the active sampling strategies (PU, AU, BADGE) drop in performance compared to random sampling, which may be due to their prioritisation of up to 10.9% more noisy samples.

2) Ablation Studies: We investigated the contributions of the noise-robust learning and evidential learning components in NR-eNoT, having covered the SSL and AL components in Subsection A. We focused these experiments on the 17.2% noise rate scenario and provide results in Table III.

a) The small-loss inclusion mechanism: Across labelling budgets, NR-eNoT improves AUPRC over eNoT by up to 2.6%. Thus, including small-loss samples during backpropagation improves robustness to noisy annotations.

b) Evidential network parameterisation: We ablated EDL in BADGE only as PU and AU rely on the evidential networks. NR-eNoT+BADGE improves AUPRC by up to 3.6% over NR-NoT+BADGE across labelling budgets. Therefore, the evidential loss enhances identification of samples with noisy annotations and mitigates error propagation compared to NR-NoT's cross-entropy loss.

C. CSEAL-MEAP with the dwMRI data annotation task

For CSEAL-MEAP, we characterise the deep learning model performance at different labelling budgets using the model trainer to assess how CSEAL impacts annotation requirements, accuracy of auto-labelling, and robustness to labelling errors during a real-world dwMRI annotation task.

1) Reducing the annotation workload: Fig. 5(a) shows the test AUPRC as a function of the labelling budget for annotation workflows A–C with labels from senior annotators. Performance under deep Bayesian active learning method B(SUP+EU) is sometimes, but not consistently, higher

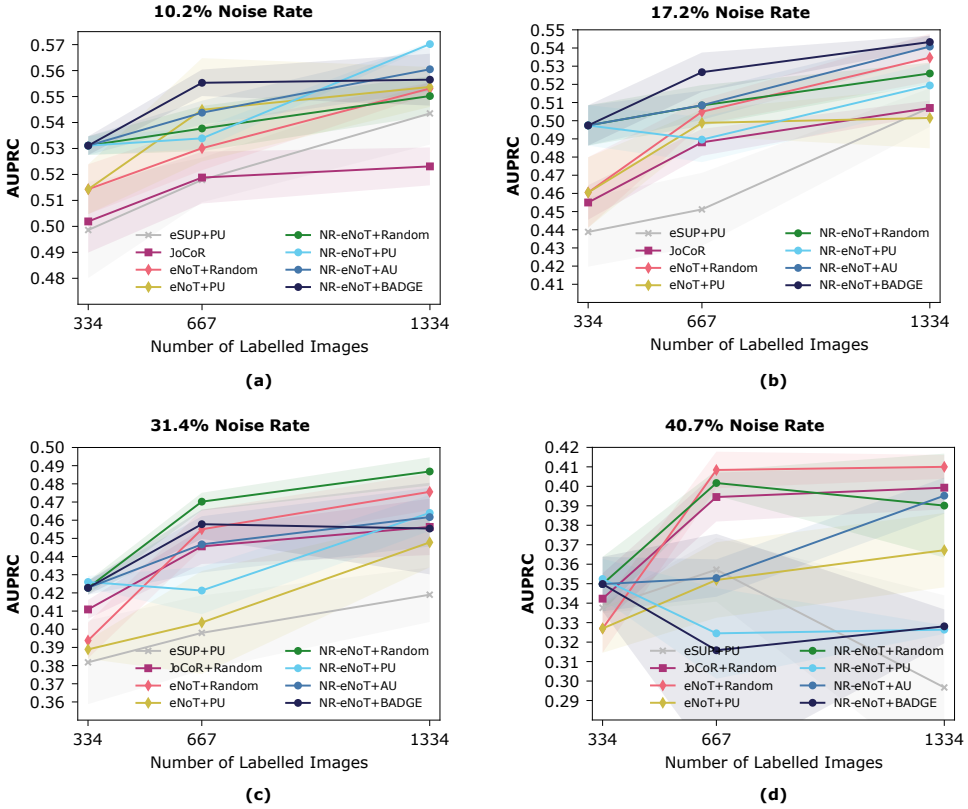

Fig. 4. Performance results (mean ± standard deviation) for image classification as a function of labelling budget on the NoisyCXR dataset. The AUPRC for noise rates of (a) 10.2%, (b) 17.2%, (c) 31.4%, and (d) 40.7%. The proposed Noise Robust-evidential NoTeacher (NR-eNoT) approach is combined with predictive uncertainty (PU), aleatoric uncertainty (AU) and BADGE for active learning as well as random sampling. It is compared with four baselines: the evidential NoTeacher (eNoT) with random and PU sampling, evidential Supervised (eSUP) and Joint Training with Co-Regularization (JoCoR).

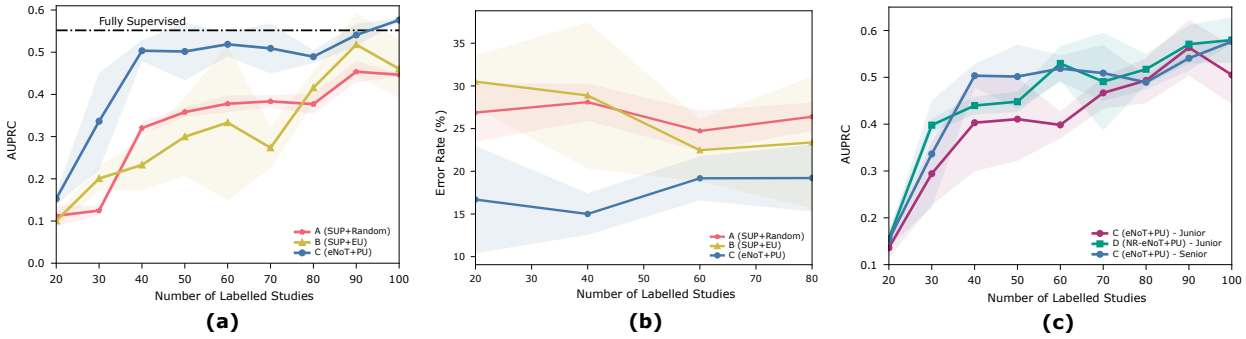

Fig. 5. Performance results of the CSEAL-assisted Medical Expert Annotation Platform (CSEAL-MEAP) on clinical brain dwMRI images. Workflow A(SUP+Random) represents a human-only paradigm while B(SUP+EU), C(eNoT+PU) and D(NR-eNoT+PU) represent those assisted with Bayesian active learning, CSEAL semi-supervised active learning, and CSEAL noise-robust SSAL, respectively. (a) Test AUPRC with senior annotators' labels in workflows A–C. (b) Percentage of studies where model-generated labels differ from the senior annotators' labels for batches of 20 auto-labelled studies in workflows A–C. (c) A comparison of the test AUPRC when models are trained with junior and senior annotators' labels in workflows C and D. All results are plotted against different labelling budgets with the mean ± standard deviation.

than that under the typical human-only supervised learning paradigm A(SUP+Random) (e.g., A has 4.5%–11.0% higher AUPRC than B for 40–70 labelled studies). Thus, conventional AL in B does not reliably reduce the annotation workload to reach the fully supervised performance. Meanwhile, performance under CSEAL method C(eNoT+PU) quickly reaches "elbow point" with only 40 labelled studies, after which it starts to plateau close to the fully supervised performance. Hence, our CSEAL approach, eNoT+PU, requires substantially fewer manually labelled studies than typical human-only or supervised paradigms. Gains of C(eNoT+PU) over B(SUP+EU) may stem from better prioritisation of abnormal studies for annotation (7.5%–11.1% more abnormal studies prioritised under workflow C than B).

2) Accuracy of auto-labelling: Fig. 5(b) shows the proportions of studies where automated labels generated by the model differ from senior annotators' labels under annotation workflows A–C. Model predictions under workflow

TABLE III
ABLATION STUDIES FOR NR-ENOT COMBINED WITH PU, AU AND BADGE UNDER THE HIGH (17.2%) NOISE RATE SCENARIO IN THE NOISYCXR DATASET. THE PARTS ABLATED ARE NOISE-ROBUST LEARNING (ENOT) AND EVIDENTIAL DEEP LEARNING (NR-NOT). THE TEST AUPRC IS REPORTED AT DIFFERENT LABELLING BUDGETS.

Method	AUPRC Number of Labelled Images (in %) 667 (5.0)	1334 (10.0)
eNoT+PU	49.88 ± 0.73	50.15 ± 1.67
eNoT+AU	48.40 ± 1.80	51.51 ± 0.97
eNoT+BADGE	50.13 ± 1.00	52.65 ± 1.49
NR-NoT+BADGE	49.06 ± 0.58	51.78 ± 0.72
NR-eNoT+PU	48.96 ± 1.24	51.94 ± 1.20
NR-eNoT+AU	50.85 ± 0.80	54.08 ± 0.61
NR-eNoT+BADGE	$\mathbf{52.67 \pm 1.08}$	$\mathbf{54.33 \pm 0.35}$

C(eNoT+PU) have 3.3%–13.9% higher accuracy than workflows A(SUP+Random) and B(SUP+EU). These results show that CSEAL-MEAP offers more accurate model-generated labels that would require fewer corrections from reviewers during annotation, hence speeding up annotation.

3) Junior versus senior annotators' labels: Fig. 5(c) shows the test AUPRC for annotation workflows C and D with labels from senior (clean) and junior (noisy) annotators. Under workflow C(eNoT+PU) with noisy annotations, AUPRC is up to 13.1% lower than with clean annotations. However, running workflow D(NR-eNoT+PU) with only noisy annotations achieves performance close to that of clean annotations (especially for 20–70 labelled studies), hence demonstrating how CSEAL mitigates the adverse impact of noisy annotations.

VI. DISCUSSION AND CONCLUSION

To address the challenges of learning with limited and noisy annotations, we introduced a principled SSAL framework named Consistency-based Semi-supervised Active Learning (CSEAL). CSEAL integrates predictive uncertainty estimation with semi-supervised active classification. Specifically, it uses evidential learning to model distributions over prediction probabilities, enforce consistency, and leverage uncertainty for active learning. CSEAL also includes enhancements for noise-robust learning. We provide customisations of CSEAL to multiple SSL and AL approaches, and perform extensive experiments on X-Ray, CT, and MRI datasets to demonstrate substantial performance gains of CSEAL over SOTA baselines. We further embed CSEAL within an assisted annotation process and demonstrate its ability to boost annotation efficiency for a real-world radiology image annotation task.

Methodologically, CSEAL advances SSAL methods [4], [6], [7] by reliably estimating predictive uncertainty using both labelled and unlabelled data through evidential deep learning. Our novel loss formulation improves generalisability across X-Ray, CT, and MRI modalities, and increases robustness to noisy annotations through effective learning of evidences. Compared to SSL [3] and AL [2] methods, CSEAL mitigates challenges of error propagation, cold-start, and uncertainty estimation. The evidential framework enables epistemic uncertainty estimation, which can be leveraged to identify out-of-distribution and uncertain samples, and improve performance on low-prevalence classes [54]. Although NR-eNoT indirectly benefits from peer learning similarly to JoCoR [48], it uniquely incorporates evidential losses and unlabelled samples to better address label noise. Unlike [42], [55], NR-eNoT addresses annotation errors without relying on availability of clean annotations, making it highly suited for practical scenarios. Moreover, we adopt a 2D CNN with scan-level aggregation to balance computational efficiency and data efficiency in low-label, iterative active learning setting. Although 3D CNNs can capture richer volumetric context, they are less practical under these constraints due to their higher computational cost and data requirements. Our aggregation strategy enables robust scan-level representations and uncertainty estimates, reduces slice-level redundancy, and supports stable and consistent sample selection across AL rounds. Importantly, CSEAL offers a general methodology within a unified framework that can accommodate varied SSL and AL approaches

Our work has demonstrated implications for radiology image annotation workflows in a proof-of-concept study. By coupling learning and annotation, CSEAL enables annotators to effectively leverage information from large unlabelled datasets in a prioritised manner. With few labelled samples (e.g., < 50 brain MRI scans), our CSEAL-assisted platform can generate accurate automated labels and enable greater annotation efficiency. With its inbuilt noise robustness mechanism, our framework can cater to both non-expert and expert annotators. Future work could explore the choice of active learning strategies to tune the performance of CSEAL, especially at very high labelling budgets scenarios with high effective noise rates and broader validation studies. Further extensions for the realm of active barely supervised learning to boost performance in settings with sparse or cross-annotation [56], incorporation of continual fine-tuning strategies [57] to improve scalability and efficiency, self-supervised pre-training approaches [58] to boost performance in early annotation rounds, and extensions to broader image analysis tasks [59], [60] could be studied.

VII. ACKNOWLEDGEMENT

We thank Dr Mike Shou from the National University of Singapore for his support and insightful discussions.

VIII. REFERENCES

[1] J. A. Dunnmon, D. Yi, C. P. Langlotz, C. Ré, D. L. Rubin, and M. P. Lungren, "Assessment of convolutional neural networks for automated classification of chest radiographs," *Radiology*, 290(2), 537–544, 2019.
[2] P. Ren *et al.*, "A survey of deep active learning," *ACM CSUR*, 54(9), 1–40, 2021.
[3] X. Yang, Z. Song, I. King, and Z. Xu, "A survey on deep semi-supervised learning," *IEEE TKDE*, 35(9), 8934–8954, 2022.
[4] M. Gao, Z. Zhang, G. Yu, S. Ö. Arık, L. S. Davis, and T. Pfister, "Consistency-based semi-supervised active learning: Towards minimizing labeling cost," in *ECCV*. Springer, 12355, 510–526, 2020.
[5] A. Oliver, A. Odena, C. A. Raffel, E. D. Cubuk, and I. Goodfellow, "Realistic evaluation of deep semi-supervised learning algorithms," in *NeurIPS*, 31, 3239–3250, 2018.
[6] S. Huang, T. Wang, H. Xiong, J. Huan, and D. Dou, "Semi-supervised active learning with temporal output discrepancy," in *ICCV*, 3447–3456, 2021.

[7] W. Zhang *et al.*, "BoostMIS: Boosting medical image semi-supervised learning with adaptive pseudo labeling and informative active annotation," in *CVPR*, 20666–20676, 2022.

[8] J. Van Amersfoort, L. Smith, Y. W. Teh, and Y. Gal, "Uncertainty estimation using a single deep deterministic neural network," in *ICML*. PMLR, 9690–9700, 2020.

[9] M. Abdalla and B. Fine, "Hurdles to artificial intelligence deployment: Noise in schemas and "gold" labels," *Radiology: Artificial Intelligence*, 5(2), e220056, 2023.

[10] C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals, "Understanding deep learning (still) requires rethinking generalization," *Communications of the ACM*, 64(3), 107–115, 2021.

[11] D. Azcona, K. McGuinness, and A. F. Smeaton, "A comparative study of existing and new deep learning methods for detecting knee injuries using the MRNet dataset," in *IDSTA*. IEEE, 149–155, 2020.

[12] W. Li *et al.*, "PathAL: An active learning framework for histopathology image analysis," *IEEE TMI*, 41(5), 1176–1187, 2021.

[13] J. Li, R. Socher, and S. C. H. Hoi, "DivideMix: Learning with noisy labels as semi-supervised learning," in *ICLR*. OpenReview.net, 2020.

[14] M. Sensoy, L. Kaplan, and M. Kandemir, "Evidential deep learning to quantify classification uncertainty," in *NeurIPS*, 31, 3183–3193, 2018.

[15] 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 *2097–2106, CVPR*, 2017.

[16] A. E. Flanders *et al.*, "Construction of a machine learning dataset through collaboration: the RSNA 2019 brain CT hemorrhage challenge," *Radiology: Artificial Intelligence*, 2(3), e190211, 2020.

[17] N. Bien *et al.*, "Deep-learning-assisted diagnosis for knee magnetic resonance imaging: development and retrospective validation of MRNet," *PLoS Medicine*, 15(11), e1002699, 2018.

[18] M. Bernhardt *et al.*, "Active label cleaning for improved dataset quality under resource constraints," *Nature Communications*, 13(1), 1161, 2022.

[19] S. Balaram, C. M. Nguyen, A. Kassim, and P. Krishnaswamy, "Consistency-based semi-supervised evidential active learning for diagnostic radiograph classification," in *MICCAI*, Springer, 675–685, 2022.

[20] B. Settles, "Active learning literature survey," 2009.

[21] O. Sener and S. Savarese, "Active learning for convolutional neural networks: A core-set approach," in *ICLR*. OpenReview.net, 2018.

[22] J. T. Ash, C. Zhang, A. Krishnamurthy, J. Langford, and A. Agarwal, "Deep batch active learning by diverse, uncertain gradient lower bounds," in *ICLR*, 2020.

[23] F. Wang, Z. Han, Z. Zhang, and Y. Yin, "Active source free domain adaptation," *arXiv preprint arXiv:2205.10711*, 2022.

[24] Y. Gal, R. Islam, and Z. Ghahramani, "Deep bayesian active learning with image data," in *ICML*. PMLR, 70, 1183–1192, 2017.

[25] F. C. Ghesu *et al.*, "Quantifying and leveraging predictive uncertainty for medical image assessment," *Medical Image Analysis*, 68, 101855, 2021.

[26] D.-H. Lee *et al.*, "Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks," in *Workshop on Challenges in Representation Learning, ICML*. Atlanta, 3(2), 896, 2013.

[27] E. Arazo, D. Ortego, P. Albert, N. E. O'Connor, and K. McGuinness, "Pseudo-labeling and confirmation bias in deep semi-supervised learning," in *IJCNN*. IEEE, 1–8, 2020.

[28] T. Miyato, S.-i. Maeda, M. Koyama, and S. Ishii, "Virtual adversarial training: A regularization method for supervised and semi-supervised learning," *IEEE TPAMI*, 4(8), 1979–1993, 2018.

[29] A. Tarvainen and H. Valpola, "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results," in *NeurIPS*, 30, 2017.

[30] B. Unnikrishnan, C. Nguyen, S. Balaram, C. Li, C. S. Foo, and P. Krishnaswamy, "Semi-supervised classification of radiology images with noteacher: A teacher that is not mean," *Medical Image Analysis*, 73, 102148, 2021.

[31] Z. Ke, D. Wang, Q. Yan, J. Ren, and R. W. Lau, "Dual student: Breaking the limits of the teacher in semi-supervised learning," in *ICCV*, 6728–6736, 2019.

[32] K. Wang, D. Zhang, Y. Li, R. Zhang, and L. Lin, "Cost-effective active learning for deep image classification," *IEEE TCSVT*, 27(12), 2591–2600, 2016.

[33] J. Guo *et al.*, "Semi-supervised active learning for semi-supervised models: Exploit adversarial examples with graph-based virtual labels," in *ICCV*, 2896–2905, 2021.

[34] Z. Borsos, M. Tagliasacchi, and A. Krause, "Semi-supervised batch active learning via bilevel optimization," in *IEEE ICASSP*, 3495–3499, 2021.

[35] L. Zhong, K. Qian, X. Liao, Z. Huang, Y. Liu, S. Zhang, and G. Wang, "Unisal: Unified semi-supervised active learning for histopathological image classification," *Medical Image Analysis*, vol. 102, p. 103542, 2025.

[36] S. Shetab Boushehri, A. B. Qasim, D. Waibel, F. Schmich, and C. Marr, "Systematic comparison of incomplete-supervision approaches for biomedical image classification," in *ICANN*. Springer, 13529, 355–365, 2022.

[37] T. Yin, N. Liu, and H. Sun, "Towards cost-effective learning: A synergy of semi-supervised and active learning," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 10 163–10 172.

[38] S. Zhao, X. Zhang, L. Pan, X. Xu, and D. Pelusi, "Evidential deep active learning for semi-supervised classification," *arXiv preprint arXiv:2505.20691*, 2025.

[39] Z. Wen, O. Pizarro, and S. Williams, "Active self-semi-supervised learning for few labeled samples fast training," *arXiv preprint arXiv:2203.04560*, 2022.

[40] F. Rodrigues and F. Pereira, "Deep learning from crowds," in *AAAI*, 32(1), 2018.

[41] B. Chen *et al.*, "Combating medical label noise via robust semi-supervised contrastive learning," in *MICCAI*. Springer, 562–572, 2023.

[42] M. A. Hussain *et al.*, "Active deep learning from a noisy teacher for semi-supervised 3D image segmentation: Application to COVID-19 pneumonia infection in CT," *Computerized Medical Imaging and Graphics*, 102, 102127, 2022.

[43] M. Aljabri, M. AlAmir, M. AlGhamdi, M. Abdel-Mottaleb, and F. Collado-Mesa, "Towards a better understanding of annotation tools for medical imaging: a survey," *Multimedia Tools and Applications*, 81(18), 25877–25911, 2022.

[44] Diaz-Pinto *et al.*, "MONAI Label: A framework for AI-assisted interactive labeling of 3D medical images," *Medical Image Analysis*, 95, 103207, 2024.

[45] A. P. Dempster, "A generalization of bayesian inference," *Journal of the Royal Statistical Society: Series B (Methodological)*, 30(2), 205–232, 1968.

[46] A. Jøsang, *Subjective Logic: A formalism for reasoning under uncertainty*. Springer Publishing Company, Incorporated, 2016.

[47] B. Han *et al.*, "Co-teaching: Robust training of deep neural networks with extremely noisy labels," in *NeurIPS*, vol. 31, 2018.

[48] H. Wei, L. Feng, X. Chen, and B. An, "Combating noisy labels by agreement: A joint training method with co-regularization," in *CVPR*, 13726–13735, 2020.

[49] C. Xu, D. Tao, and C. Xu, "A survey on multi-view learning," *arXiv:1304.5634*, 2013.

[50] J. Zech, "reproduce-chexnet," GitHub repository, 2018. [Online]. Available: https://github.com/jrzech/reproduce-chexnet

[51] X. Wang *et al.*, "A deep learning algorithm for automatic detection and classification of acute intracranial hemorrhages in head CT scans," *NeuroImage: Clinical*, 32, 102785, 2021.

[52] Y. Yu *et al.*, "A multitask framework for label refinement and lesion segmentation in clinical brain imaging," in *MILLanD*. Springer, 60–70, 2023.

[53] P. Rajpurkar *et al.*, "CheXNet: Radiologist-level pneumonia detection on chest X-rays with deep learning," *arXiv:1711.05225*, 2017.

[54] C. Qin, Y. Wang, F. Zeng, J. Zhang, Y. Cao, X. Yin, S. Huang, D. Chen, H. Zhang, and Z. Ju, "Active domain adaptation based on probabilistic fuzzy c-means clustering for pancreatic tumor segmentation," *IEEE Transactions on Fuzzy Systems*, 2025.

[55] Y. Wen, "Active semi-supervised learning based on global uncertainty variation with noise resistance," in *IEEE CAI*. IEEE, 89–95, 2024.

[56] X. Wu, Z. Xu, and R. K.-y. Tong, "Few slices suffice: Multi-faceted consistency learning with active cross-annotation for barely-supervised 3d medical image segmentation," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 2024, pp. 286–296.

[57] Z. Zhou, J. Y. Shin, S. R. Gurudu, M. B. Gotway, and J. Liang, "Active, continual fine tuning of convolutional neural networks for reducing annotation efforts," *Medical image analysis*, 71, 101997, 2021.

[58] Y. He *et al.*, "Geometric visual similarity learning in 3D medical image self-supervised pre-training," in *CVPR*, 9538–9547, 2023.

[59] M. Liu *et al.*, "A deep learning method for breast cancer classification in the pathology images," *IEEE J-BHI*, 26(10), 5025–5032, 2022.

[60] S. Graham *et al.*, "One model is all you need: multi-task learning enables simultaneous histology image segmentation and classification," *Medical Image Analysis*, 83, 102685, 2023.