Uncertainty-aware Health Diagnostics via Class-balanced Evidential Deep Learning

Tong Xia, Ting Dang Member, IEEE, Jing Han Member, IEEE, Lorena Qendro, Cecilia Mascolo

Abstract— Uncertainty quantification is critical for ensuring the safety of deep learning-enabled health diagnostics, as it helps the model account for unknown factors and reduces the risk of misdiagnosis. However, existing uncertainty quantification studies often overlook the significant issue of class imbalance, which is common in medical data. In this paper, we propose a class-balanced evidential deep learning framework to achieve fair and reliable uncertainty estimates for health diagnostic models. This framework advances the state-of-the-art uncertainty quantification method of evidential deep learning with two novel mechanisms to address the challenges posed by class imbalance. Specifically, we introduce a pooling loss that enables the model to learn less biased evidence among classes and a learnable prior to regularize the posterior distribution that accounts for the quality of uncertainty estimates. Extensive experiments using benchmark data with varying degrees of imbalance and various naturally imbalanced health data demonstrate the effectiveness and superiority of our method. Our work pushes the envelope of uncertainty quantification from theoretical studies to realistic healthcare application scenarios. By enhancing uncertainty estimation for class-imbalanced data, we contribute to the development of more reliable and practical deep learning-enabled health diagnostic systems.[1]

Index Terms— Uncertainty quantification, deep learning, trustworthy health diagnostics, class imbalance

I. Introduction

Deep learning has demonstrated impressive performance across various domains, but its lack of interpretability due to black-box models has hindered its trustworthiness. Uncertainty quantification plays a vital role in addressing this concern by allowing deep neural networks to recognize and communicate their level of confidence. Uncertain-aware models are credible as they are aware of what is known and what is unknown. This capability is particularly crucial in safety-critical applications like health diagnostics [1], [2]. As depicted in Fig. 1, defensibly quantifying uncertainty for automatic diagnostics, if achieved, will greatly enhance the reliability of deep learning in real-world medical applications.

The manuscript is submitted in June 2023. This work was supported by ERC Project 833296 (EAR).

Tong Xia, Jing Han, and Cecilia Mascolo are with the Department of Computer Science and Technology, University of Cambridge. Ting Dang and Lorena Qendro are with Nokia Bell Labs (Cambridge), and Ting is also affiliated with the University of Cambridge and the University of New South Wales.

Corresponding to Tong Xia (tx229@cam.ac.uk).

[1]Code available at https://github.com/XTxiatong/Class-balanced-EDL.git

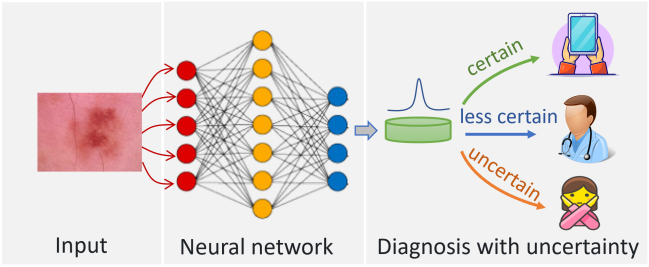


Fig. 1. An uncertainty-aware deep learning driven health diagnostics system. This system aims to provide uncertainty estimation alongside the diagnosis for each instance. A low uncertainty value indicates a confident automatic diagnosis where the model outputs can be trustworthy. However, if the prediction exceeds the model's capabilities and the uncertainty is relatively high, the instance will be referred to doctors for clinical confirmation. Furthermore, instances with extremely high uncertainty are likely to stem from data sources outside the training data distribution, such as variations in data collection devices, data corruption, or unseen pathology not encountered during training. In such cases, the system will ignore the instances or require data recollection.

Health diagnostics inherently involves classification tasks, where deep neural networks extract features from the inputted medical data to predict the most likely associated disease category. In the past decade, various approaches have been proposed to quantify the classification uncertainty [3], [4], and some methods have been explored on specific medical datasets [5]–[10]. For instance, Pacheco et al. compared several uncertainty estimation methods on skin lesion images to detect the out-of-distribution inputs [9], and Kang et al. explored statistical uncertainty entropy to identify noisy physiologic signals that the model classifies incorrectly [10]. However, there is a notable lack of comprehensive studies that evaluate the robustness and generalizability of these approaches across various medical modalities and architectures. Addressing this gap is crucial to determine the practical viability and applicability of uncertainty quantification methods in diverse healthcare settings.

Furthermore, we notice that a considerable issue, class imbalance, is often overlooked in most existing works, even though it is prevalent in medical data and can significantly impact the quality of estimated uncertainties. Class imbalance refers to a substantial disparity in the number of samples between different classes. Particularly in medical datasets, due to the low prevalence rate of certain diseases, it is usually more challenging to gather enough samples from unhealthy participants compared to healthy

controls [11]. For instance, in the largest skin lesion image datasets [12], there are six pathologies covered, with the largest class accounting for 67.1% and the smallest class accounting for only 1.1% of the entire dataset. The limited number of samples in the minority classes, representing the less prevalent unhealthy categories, poses a challenge for the model to accurately capture the underlying patterns and variances associated with those disease categories. Consequently, the uncertainty estimates for the minority classes may be mis-estimated, leading to an erroneous sense of confidence in the diagnoses. To the best of our knowledge, there is limited research that considers the impact of data imbalance when estimating uncertainty for deep health diagnostics. Xia et al. proposed a method that involves re-sampling the data to generate balanced data bags for ensemble learning, resulting in accurate predictions and high-quality uncertainty estimation for the binary COVID-19 detection problem [13]. However, this method is infeasible and inefficient for multi-class health diagnostics. In light of this, this paper aims to address the challenge of uncertainty-aware modelling for various class imbalances to improve the accuracy and reliability of deep learning-driven health diagnostics.

Considering that among various uncertainty quantification methods, the recently emerged evidential deep learning (EDL) becomes a standout due to its impressive efficiency and effectiveness [14]–[17]. The core principle of EDL is to learn the evidence for classification, which will be mapped into a Dirichlet distribution over the categorical prediction for uncertainty quantification. As EDL solely modifies the output of a classification model without altering the feature-extracting architecture, it also offers the advantage of leveraging pre-trained models to quantify uncertainty, particularly in scenarios with limited medical data availability [18]. All those properties make EDL a promising solution for health diagnostic uncertainty quantification, and herein in our study, we will focus on EDL approaches.

Despite showing great promise, most investigations and evaluations of EDL rely on well-curated datasets and balanced machine learning benchmarks such as CIFAR10 and CIFAR100 [17], [19], [20], leaving the impact of medical data unclear. To gain deeper insights into the behaviour of EDL, we conducted an analysis using an imbalanced dataset to identify potential confounding factors. This revealed a negative association with the quality of uncertainty measurements. Consequently, our findings indicate that EDL would also be negatively affected by class imbalance mainly for two reasons: i) the uniform empirical loss across all samples could introduce classification bias, and ii) EDL assumes a uniform distribution across all classes, which does not reflect the real data distribution. Consequently, with severe class imbalance, EDL may produce imprecise Dirichlet distributions, resulting in low uncertainty quality. This can lead to confident but incorrect diagnoses, while inadequate uncertainty quantification prevents clinicians from correcting automatic diagnoses.

The analysis emphasizes the importance of addressing the limitations posed by the uniform empirical loss and uniform prior distribution loss in improving EDL on imbalanced medical data. This paper presents two novel mechanisms to tackle this problem. Firstly, we propose a customized class-level pooling loss to alleviate bias in the classification evidence. Additionally, we advocate for the adoption of a learnable prior that is regulated by the class distribution, thereby enhancing the learning process for minority classes. Through comprehensive experiments, we demonstrate that our proposed approach achieves fair uncertainty estimation for all classes, thereby paving the way for more reliable automatic diagnostic systems.

The main contributions of this paper are summarized below,

- We shift the attention of uncertainty quantification from well-curated data to real-world health data with skewed class distribution, and we provide a systematic uncertainty study on various health diagnostics scenarios.
- To address the challenges caused by class imbalance, this paper proposes a class-balanced EDL framework. This framework incorporates the start-of-the-art uncertainty quantification method EDL with two novel mechanisms to produce fair and reliable uncertainty estimates for all health conditions.
- We evaluate our class-balanced EDL framework on three real-world health diagnostic tasks with different data modalities and model architectures. Our method presents superior performance against the state-of-the-art baselines for class imbalance and uncertainty quantification. Particularly, we improve the accuracy of misdiagnosis identification and out-of-training-distribution detection by up to 16.1%.

II. Preliminary

A. Problem Definition

In this paper, we target diagnostic classification tasks: assuming a training dataset $\mathcal{D} = \{X^{(i)}, y^{(i)}\}_{i=1}^{N}$ is available, where $X^{(i)}$ denotes the input, $y^{(i)}$ corresponds to a disease type among C total categories and N is the number of training samples. N_c presents the number of samples for class c. We consider a skewed training distribution and thus, N_c varies among classes. We term class c a majority class if $N_c > N/C$, otherwise a minority class. The task is to learn a neural network parameterized by θ that predicts $\hat{y}^{(i)}$ with an uncertainty measurement $\hat{u}^{(i)}$ for any given sample $X^{(i)}$ from the testing set.

B. Limitation of Softmax Probability

Despite the prevalence of estimating uncertainty directly from the Softmax layer in deep learning [21], this approach exhibits significant limitations when compared to more advanced uncertainty estimation methods such as EDL. Specifically, softmax layer aims to transfer the logits $\boldsymbol{z}^{(i)}$ into categorical probabilities $\boldsymbol{p}^{(i)} = [p_1^{(i)}, p_2^{(i)}, ..., p_C^{(i)}]$, where $p_c^{(i)} = e^{z_c^{(i)}} / \sum_{j=1}^{C} e^{z_j^{(i)}}$ and $\sum_{c=1}^{C} p_c^{(i)} = 1$, for

a given input $X^{(i)}$. The predicted entropy of $\boldsymbol{p}^{(i)}$ as formulated by, $Entropy(\boldsymbol{p}^{(i)}) = \sum_{c=1}^{C} -p_c^{(i)} \cdot \log(p_c^{(i)})$, is commonly used as the uncertainty measurement, as it indicates the confidence of the prediction. However, $\boldsymbol{p}^{(i)}$ is a point estimation that cannot capture the uncertainty inheriting from the model [22]. Besides, the probability yielded by softmax is usually over-confident as it always predicts a close set for any given inputs, while the real world is open with unseen classes [23]. Polished English: For instance, the uncertainty of a Softmax model, previously trained to distinguish between individuals with Asthma and healthy individuals, may not be sufficient to identify a new disease beyond Asthma when it occurs.

C. Evidential Deep Learning

Different from Softmax, EDL leverages Dirichlet distribution $\boldsymbol{q}^{(i)}$ – the distribution over $\boldsymbol{p}^{(i)}$, to achieve prediction and uncertainty quantification simultaneously [24], [25]. The Dirichlet distribution is used because it is the natural conjugate posterior of multinomial distribution (i.e., the probability $\boldsymbol{p}^{(i)}$ can be regarded as a multinomial distribution). Underpinned by the Bayesian rule, EDL aims to capture the classification evidence $\boldsymbol{l}^{(i)}$ by the deep learning model and then transform a uniform prior $Dir(1)$ into the posterior $\boldsymbol{q}^{(i)} = Dir(\boldsymbol{\alpha}^{(i)})$, with $\boldsymbol{\alpha}^{(i)} = 1 + \boldsymbol{l}^{(i)}$ [25]. More specifically, the posterior $\boldsymbol{q}^{(i)} = Dir(\boldsymbol{\alpha}^{(i)})$ is parameterized by $\boldsymbol{\alpha}^{(i)} = [\alpha_1^{(i)}, \alpha_2^{(i)}, ..., \alpha_C^{(i)}]$ for C classes, where $\alpha_c^{(i)} = 1 + l_c^{(i)}$.

The posterior Dirichlet distribution can be viewed as an infinite ensemble of point estimations $\boldsymbol{p}^{(i)}$. Therefore, EDL enables a better-calibrated way of quantifying epistemic uncertainty compared to traditional softmax-based deep learning [14], [15]. Additionally, the expectation of probability $\hat{\boldsymbol{p}}^{(i)}$ presents the average predictive confidence which reflects the aleatoric uncertainty. EDL is also able to capture the distributional shift: if no remarkable evidence can be modelled for a given input, the posterior $\alpha_c, \forall c \in C$ will approach 1, i,e., the prior. Overall, given an input $X^{(i)}$, an EDL model f_θ outputs distribution $\boldsymbol{q^{(i)}} = Dir(\boldsymbol{\alpha}^{(i)})$ with the predictive probability $\hat{\boldsymbol{p}}^{(i)}$, categorical prediction $\hat{y}^{(i)}$ and uncertainty measurement of Differential Entropy ($DE^{(i)}$) inferred as below,

$$
\begin{aligned}
&\boldsymbol{\alpha}^{(i)} = 1 + \boldsymbol{l}^{(i)}, \\
&\hat{p}_c^{(i)} = \mathbb{E}[p_c^{(i)}] = \frac{\alpha_c^{(i)}}{\alpha_0^{(i)}}, \\
&\hat{y}^{(i)} = \arg\max_c \mathbb{E}[p_c^{(i)}], \\
&DE^{(i)} = \mathbb{E}_{\boldsymbol{p}^{(i)} \sim \boldsymbol{q}^{(i)}}[Entropy(\boldsymbol{p}^{(i)})],
\end{aligned}
\tag{1}
$$

where $\alpha_0^{(i)} = \sum_{c=1}^{K} \alpha_c^{(i)}$. DE reflects how the energy is distributed, i.e., the "peakedness", in the Dirichlet distribution. A larger DE corresponds to a higher uncertainty of a prediction. Some illustrative examples of the posterior Dirichlet distributions are given in Fig. 2.

EDL can be adapted to any neural network architecture by simply replacing the softmax layer with a plunge-in Dirichlet distribution estimation layer on the output side [14], [15]. Following the latest literature [16], [26], the following objective is used to optimize θ,

$$
\begin{aligned}
\min_{\theta} \quad & \mathcal{L} = \frac{1}{N} \sum_{i=1}^{N} \mathcal{L}^{(i)}, \\
& \mathcal{L}^{(i)} = \mathbb{E}_{\boldsymbol{p}^{(i)} \sim \boldsymbol{q}^{(i)}}[\mathcal{C}(\boldsymbol{p}^{(i)}, y^{(i)})] + \lambda \cdot \mathcal{L}_r^{(i)},
\end{aligned}
\tag{2}
$$

where $\mathcal{C}$ denotes the cross-entropy, i.e., $\mathcal{C}(\boldsymbol{p}^{(i)}, y^{(i)}) = -\log p_{y^{(i)}}^{(i)}$ for sample i, and $\mathcal{L}_r^{(i)} = KL[(Dir(\boldsymbol{\alpha}^{(i)})||Dir(\mathbf{1})]$ denotes a regularization for each posterior $\boldsymbol{q}^{(i)}$.

Given the above characteristics, EDL is more effective than Softmax in quantifying the uncertainty arising from both data and model, culminating the predictive probabilities, and further being aware of the distributional shift.

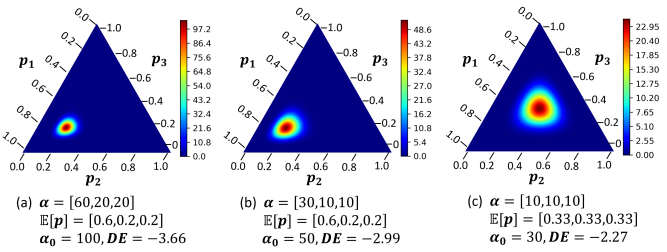


Fig. 2. Three-class Dirichlet distribution. (a) and (b) point to the same predicted class, but (a) is sharper so it is more confident while (b) is more uncertain with a larger $\boldsymbol{DE}$. (c) shows an example that is certain to none of the classes.

III. Impact of Class Imbalance on EDL

In this section, we conduct empirical and theoretical analyses of EDL to uncover its limitations in handling class imbalance presented in the data.

A. Empirical Observation

To effectively showcase the influence of class imbalance on the performance of EDL, we conducted experiments using a benchmark dataset with varying degrees of class imbalance. Specifically, we implemented an image classification task using the EDL loss (Eq. 2) and compared the results obtained with both balanced and imbalanced data through downsampling (after down-sampling, the data exhibited a skewed step distribution, mimicking class imbalance). For the experiments, we utilized the VGG model to classify the CIFAR10 dataset [27]. Specifically, VGG16 pre-trained by ImageNet is used as the backbone model, followed by a linear layer to output the parameters α to formulate the Dirichlet distribution. The data distribution and results of our experiments are presented in Fig. 3.

As it can be observed in Fig. 3(a), with balanced training data, for all 10 classes, the EDL quantifies higher uncertainty for incorrect predictions than correct predictions within each class. This suggests that quantified uncertainty can reliably reflect the confidence of the model. However, this no longer holds with a skewed distribution as displayed in Fig. 3(b): incorrect predictions from minority classes, e.g., class 1 and 2, manifest very low uncertainty. This evidence verifies our concern that the EDL is vulnerable in the class imbalance scenario, and a comprehensive comprehension of the underlying principles

and viable solutions should be proposed to enhance EDL for imbalanced datasets.

B. Theoretical Analysis

In addition to empirical analysis, we also draw on theoretical insights to systematically explain the reasons behind the failure of EDL in the presence of class imbalance.

Lemma I. The across-sample empirical loss Eq. 2 induces the bias in EDL.

Given C classes, the objective in Eq. 2 can be rewritten as,

$$\begin{aligned}\min_{\theta} \quad \mathcal{L} &= \frac{1}{N}\sum_{c=1}^{C}\sum_{y^{(i)}\in c}\mathcal{L}^{(i)} \\ &= \sum_{c=1}^{C}\frac{N_c}{N}\cdot\frac{1}{N_c}\sum_{y^{(i)}\in c}\mathcal{L}^{(i)} \\ &= \sum_{c=1}^{C}\frac{N_c}{N}\cdot\overline{\mathcal{L}_c},\end{aligned} \tag{3}$$

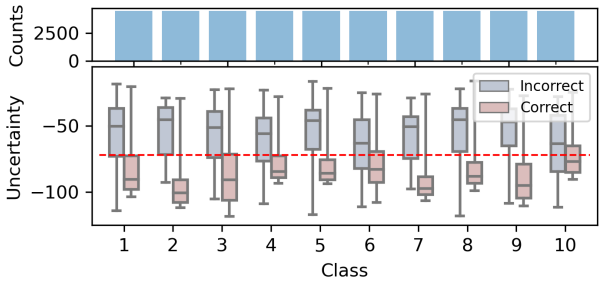
where $\overline{\mathcal{L}_c}$ presents the average loss for class c. It can be noted that class-averaged loss $\overline{\mathcal{L}_c}$ is weighted by the proportion of the samples in the training set. Herein, the object tends to prioritize optimizing $\overline{\mathcal{L}_c}$ for the majority classes. Because of the relatively small N_c, misclassifications or over-confident posteriors from minority classes could be under-looked, leading to imprecise estimation of classification evidence $\boldsymbol{l}$ (see Eq. 1). Particularly, when N_c for minority classes is extremely small, which is common for many realistic health applications where rare classes exist, the learned evidence can be more biased. As a consequence, the quantified uncertainty parameterized by $\boldsymbol{\alpha}$ could be less reliable for the minority classes due to the lack of training data.

Lemma II. The uniform prior is not feasible for EDL in the presence of imbalanced data.

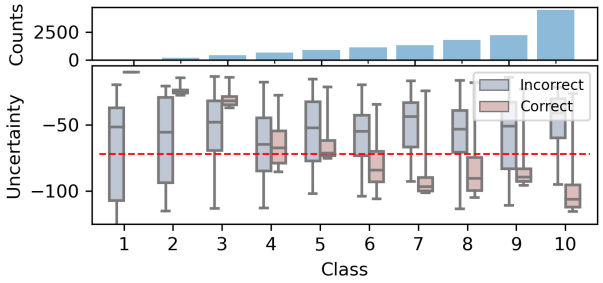
From Sec. II-C, EDL assumes a uniform $Dir(1)$ as a prior, which indicates an equal likelihood for all classes if the same amount of evidence has been observed. The regularization of the posterior, i.e., $\mathcal{L}_r^{(i)}$ in Eq. 2, also imposes a uniform smoothing across all classes, ignoring the varied learning difficulty among classes. This may not be optimal in the presence of imbalanced data, particularly when the minority classes are underrepresented with a few samples. In traditional softmax-based deep learning, classification thresholds can be adjusted (i.e., not the same threshold for every class) to allow some marginal samples to be classified into minority classes [28], [29]. Similarly, finding a suitable prior that can better regularize the posterior can be helpful in EDL.

IV. Our Method

Built upon the above discussions, we now present our novel method to enable EDL for imbalanced data. Our efforts include two aspects: (1) learning less biased evidence and (2) seeking a better prior, which are introduced below,



(a) Balanced data yields reliable uncertainty estimates.



(b) Imbalanced training data leads to poor uncertainty estimates for the minority classes.

Fig. 3. Uncertainty quantified by EDL for CIFAR10 classification. The top subfigures present the training data distribution, and the bottoms show the uncertainty for correct and incorrect predictions within each class (a larger value indicates that the prediction is less certain). The red line represents an uncertainty threshold that leads to the highest accuracy in misclassification identification.

Mechanism I. Class pooling loss. As discussed in Lemma I, the uneven distribution of samples across classes is the devil that introduces bias into the model, leading to unequal learning speeds for the classes. To overcome this issue, we propose to give equal attention to all classes no matter the number of training samples. To achieve this, we leverage a class-level pooling loss that is first calculated within each class and then averaged across classes. Specifically, $\mathcal{L}'$ in Eq. 2 will be replaced by,

$$\mathcal{L}' = \frac{1}{C}\sum_{c=1}^{C}\frac{1}{N_c}\sum_{y^{(i)}\in c}\mathcal{L}^{(i)}, \tag{4}$$

where N_c is the cardinality of class c. Thereby, in contrast to Eq. 3, $\mathcal{L}'$ is class distribution agnostic. In other words, N_c/N is fixed and cannot be approached to $1/C$. Just rephrase as 'we mitigate the bias by substituting the class-dependent weight N_c/N with a uniform weight of $1/C$.

Mechanism II. Adaptive prior. Since the uniform prior assumption has limited capacity as discussed in Lemma II., we propose to replace the uniform prior with a trainable prior parameterized by $\boldsymbol{\beta} = [\beta_1, \beta_2, ..., \beta_C]$. Learning the classification evidence through the neural network usually needs more data and could be biased, but optimizing the posterior from the prior (i.e., via $\mathcal{L}_r^{(i)}$) could be more helpful, particularly when the training data is limited. A good prior should consider the class distribution of the training data, and compensate for the class skew to ease the learning of the posterior. Herein, we propose that $\boldsymbol{\beta}$ can mimic the reversed class proportion, termed by $\boldsymbol{\eta} =$

$[N/N_1, N/N_2, ...N/N_C]$ with $\eta_c = N/N_c$. Furthermore, although it is meaningful to use $\boldsymbol{\eta}$ as $\boldsymbol{\beta}$, we do not fix the value but use a trainable prior: this allows the prior with more optimization space considering the varying learning difficulty for different classes. To achieve this, another term that measures the KL-divergence between the two categorical distributions parameterized by $\boldsymbol{\beta}$ and $\boldsymbol{\eta}$, formulated by,

$$\mathcal{L}'_p = KL[Cat(\boldsymbol{\beta})||Cat(\boldsymbol{\eta})] = \sum_{c=1}^{C} \beta_c \log \frac{\beta_c}{\eta_c}, \tag{5}$$

will be added to the objective. Correspondingly, the regularization term in Eq. 2 becomes the KL-divergence between the posterior and the trainable prior to ensure "fidelity-to-prior" [30]. The prior can be jointly optimized by the loss as formulated in Eq. 4. We term the new posterior parameterized by $\boldsymbol{\alpha}'$, and the new regularization for the posterior is denoted by $\mathcal{L}'^{(i)}_r$, which can be further derived by,

$$\begin{aligned}\mathcal{L}'^{(i)}_r &= KL[Dir(\boldsymbol{\alpha}'^{(i)})||Dir(\boldsymbol{\beta})] \\ &= \int Dir(\boldsymbol{p}|\boldsymbol{\alpha}'^{(i)}) \log \frac{Dir(\boldsymbol{p}|\boldsymbol{\alpha}'^{(i)})}{Dir(\boldsymbol{p}|\boldsymbol{\beta})} d\boldsymbol{p} \\ &= \int Dir(\boldsymbol{p}|\boldsymbol{\alpha}'^{(i)}) \big(\log Dir(\boldsymbol{p}|\boldsymbol{\alpha}'^{(i)}) - \log Dir(\boldsymbol{p}|\boldsymbol{\beta}) \big) d\boldsymbol{p}.\end{aligned} \tag{6}$$

Since the integration can be derived by digamma function ψ and gamma function Γ, i.e., $\int Dir(\boldsymbol{p}|\boldsymbol{\alpha}) \log Dir(\boldsymbol{p}|\boldsymbol{\alpha}) d\boldsymbol{p} = \int Dir(\boldsymbol{p}|\boldsymbol{\alpha}) \big[\log \Gamma(\alpha_0) - \sum_{c=1}^{C} \log \Gamma(\alpha_c) + \sum_{c=1}^{C} (\alpha_c - 1) \log \boldsymbol{p} \big] d\boldsymbol{p} = \log \Gamma(\alpha_0) - \sum_{c=1}^{C} \log \Gamma(\alpha_c) + \sum_{c=1}^{C} \alpha_c (\psi(\alpha_c) - \psi(\alpha_0))$. The closed form of $\mathcal{L}'^{(i)}_r$ is written as,

$$\begin{aligned}\mathcal{L}'^{(i)}_r = \log \Gamma(\alpha'^{(i)}_0) - \sum_{c=1}^{C} \log \Gamma(\alpha'^{(i)}_c) - \log \Gamma(\beta_0) + \\ \sum_{c=1}^{C} \log \Gamma(\beta_c) + \sum_{c=1}^{C} (\alpha'^{(i)}_c - \beta_c)(\psi(\alpha'^{(i)}_c) - \psi(\alpha'^{(i)}_0))],\end{aligned} \tag{7}$$

where $\beta_0 = \sum_{c=1}^{C} \beta_c$.

Overall objective. Integrating Mechanism I. and II., with Eq. 2, 4, 5, and 7 merged into the overall optimization loss for class-balanced EDL as,

$$\begin{aligned}\min_{\theta, \boldsymbol{\beta}} \quad \mathcal{L}' &= \frac{1}{C} \sum_{c=1}^{C} \frac{1}{N_c} \sum_{y^{(i)} \in c} \mathcal{L}'^{(i)} + \mu \cdot \mathcal{L}'_p, \\ \boldsymbol{\alpha}'^{(i)} &= \boldsymbol{\beta} + \boldsymbol{l}^{(i)}, \\ \mathcal{L}'^{(i)} &= \mathbb{E}_{\boldsymbol{p}^{(i)} \sim Dir(\boldsymbol{\alpha}'^{(i)})} [\mathcal{C}(\boldsymbol{p}^{(i)}, y^{(i)})] + \lambda \cdot \mathcal{L}'^{(i)}_r, \\ \mathcal{L}'^{(i)}_r &= KL[Dir(\boldsymbol{\alpha}'^{(i)})||Dir(\boldsymbol{\beta})], \\ \mathcal{L}'_p &= KL[Cat(\boldsymbol{\beta})||Cat(\boldsymbol{\eta})],\end{aligned} \tag{8}$$

where hyper-parameters λ and μ trade off the classification, the regularization of posterior $\mathcal{L}'_r$, and the regularization of prior $\mathcal{L}'_p$. It is now easy and efficient to calculate the loss without sampling $\boldsymbol{p}^{(i)}$ from $\boldsymbol{q}^{(i)}$ to obtain the expectation, with the closed form specified by,

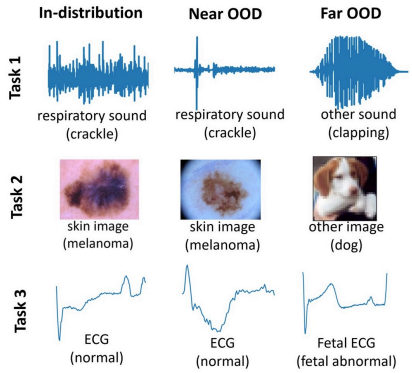


Fig. 4. Examples of the in and out-of-distribution testing samples for the three tasks.

$$\begin{aligned}\mathcal{L}' = \frac{1}{C} \sum_{c=1}^{C} \frac{1}{N_c} \sum_{y^{(i)} \in c} \{ \psi(\alpha'^{(i)}_0) - \psi(\alpha'^{(i)}_{y^{(i)}}) + \lambda \cdot [\log \Gamma(\alpha'^{(i)}_0) \\ - \sum_{c=1}^{C} \log \Gamma(\alpha'^{(i)}_c) - \log \Gamma(\beta_0) + \sum_{c=1}^{C} \log \Gamma(\beta_c) \\ + \sum_{c=1}^{C} (\alpha'^{(i)}_c - \beta_c)(\psi(\alpha'^{(i)}_c) - \psi(\alpha'^{(i)}_0))] \} + \mu \cdot \sum_{c=1}^{C} \beta_c \log \frac{\beta_c}{\eta_c},\end{aligned} \tag{9}$$

where $\alpha'^{(i)}_0 = \sum_{c=1}^{C} \alpha'^{(i)}_c$, $\beta_0 = \sum_{c=1}^{C} \beta_c$, ψ is the digamma function and Γ denotes the gamma function.

V. Experiments on Realistic Applications

To validate the effectiveness of our class-balanced EDL for real-world health diagnostics, we employ three medical tasks for experiments. The datasets used encompass various modalities, and all of them exhibit severe class imbalance, making them ideal test beds for the evaluation.

A. Datasets and Tasks

We conducted extensive experiments on three medical tasks with different data modalities. We split each dataset into a training and a testing set. The training set including a part for validation is used for model parameter learning, while the testing set is leveraged to report the performance. For each task, we also included two OOD testing sets, i.e., near OOD and far OOD. The near OOD set has the same classes as the training data but was collected with a different protocol, thus presenting a semantic shift, while the far OOD set contains similar inputs to unseen classes. The details of those datasets are elaborated below.

Task 1: Respiratory abnormity detection. We explored the potential of lung sounds for detecting respiratory abnormalities by distinguishing abnormal lung sounds from healthy sounds, leveraging the state-of-the-art ResNet34-based acoustic model [31].

- (ID) ICBHI 2017 Respiratory Challenge published a dataset collected from multiple microphones and stethoscopes [32]. The total 6,898 samples from 126 patients cover four classes: normal lung sounds (52.8%), crackle only (27.0%), wheeze only (12.9%), and both crackle and wheeze (7.3%).

TABLE I
A SUMMARY OF REAL APPLICATION DATASETS. #TRAIN IS THE ORIGINAL TRAINING DATA SIZE, WHICH IS SPLIT INTO TRAINING AND VALIDATION FOLDS WITH DIFFERENT SEEDS. #TEST IS THE TESTING SIZE. C IS THE NUMBER OF CLASSES AND D IS THE INPUT DATA DIMENSION.

Task			Dataset						OOD Dataset			
Name	Backbone	Modality	Name	#Train	#Test	C	Ratio (%)	D	Near OOD	Size	Far OOD	Size
Task 1	ResNet34	Audio	ICBHI2017	4,274	2,641	4	52.8/27.0/12.9/7.3	1×32,000	Stethoscope	336	ARCA23K	2,264
Task 2	DenseNet121	Image	HAM10000	7,206	2,809	7	67.1/11.1/11.0/5.1/3.3/1.4/1.1	3×600×450	ISIC2017	1,824	CIFAR-10	10,000
Task 3	FCNet	ECG	ECG5000	4,500	500	5	58.4/35.3/3.9/2.0/0.5	1×140	ECG200	200	FetalECG	1,965

- (Near OOD) A similar audio dataset named Stethoscope consists of 336 normal, crackle, and wheeze audio samples [33]. This dataset was collected via a 3M Littmann electronic device and thus is different from ICBHI. The demographics of this dataset and ICBHI are also different, so we used it as ICBHI's co-variate shift counterpart.
- (Far OOD) ARCA23K is a dataset of labelled sound events originating from Freesound, and each clip belongs to one of 70 typically audio classes including music, human sounds, animal sounds, etc [34]. We used the validation set containing 2,264 clips. This dataset contains different classes compared to ICBHI.

Setting. For the ID data, we followed the official patient-independent training and testing splits of the Challenge. Samples from 47 patients were used for testing, while for the rest of the patients, we randomly divided them into five folds and held out one fold per run to conduct five-fold cross-validation. For all ID and OOD datasets, audio recordings were re-sampled to 4KHz and divided into 8s clips. The clips were then transformed into Mel-spectrograms as the inputs of the model.

Task 2: Skin lesion screening. The classification of skin lesions was examined using an image classification model based on DenseNet121 [9].

- (ID) HAM10000 contains 10,015 dermatoscopic skin tumour images taken from multiple devices and demographics [12]. The image size is 600×450. The skin condition is labelled as one of the following classes: melanocytic nevi (67.1%), melanoma (11.1%), benign keratosis-like lesion (11.0%), basal cell carcinoma (5.1%), actinic keratoses (3.3%), vascular lesion (1.4%), or dermatofibroma (1.1%).
- (Near OOD) Another skin lesion dataset with 2,000 high-resolution varied-size images published by ISIC 2017 was used [35]. It was collected by another institute with a varied device from HAM10000, therefore we regard it as the near OOD.
- (Far OOD) The image classification benchmark CIFAR-10 with 10 non-skin classes was utilized as the far OOD. We used this data to simulate the scenario when a non-clinician image is input into the model.

Setting. For ID data, 30% was held out as the testing set, and five-fold cross-validation was implemented: four-fifths of the remaining 70% of the data for training and one-fifth for validation per running. Images in ISIC2017 datasets were resized uniformly to 767×1,022 before feeding into the model.

Task 3: Heart failure prediction. The detection of cardiovascular diseases was investigated using electrocardiogram (ECG) data with the one-dimensional convolutional neural network FCNet [36].

- (ID) ECG5000 is a 20-hour long one-channel ECG dataset, which was split and interpolated into equal-length (140) heart beats [37]. It consists of five classes: 58.4% are normal, 35.3% have heat failure typed R-on-T phenomenon, 3.9% PVC (Premature Ventricular Contraction), 2.0% ST (ST Segment Elevation), and 0.5% UB (Upright Biphasic T-wave).
- (Near OOD) Another dataset consisting of 200 ECG recordings with a length of 178 was used as the near OOD. This data was acquired through a method different from ECG5000 and contains two classes of normal heartbeats and Myocardial Infarction [38].
- (Far OOD) A non-invasive fetal ECG dataset consists of 1,965 heartbeats with a length of 750 [39]. Sensors were positioned on the mother's abdomen to detect and record the electrical signals produced by the fetal heart. Fetal ECG typically exhibits lower amplitude compared to that of adults, making it suitable as the far ODD dataset in our study.

Setting. We utilized a subset of 500 samples in the ID ECG5000 datasets for testing and split the rest into five folds uniformly for cross-validation.

For the aforementioned three tasks, the used datasets and model backbones are summarized in Table I. Examples of the in and out-of-distribution testing samples are given in Fig. 4.

B. Baselines

The backbone model with Softmax probability, termed as Vanilla, is implemented for each task as a basic baseline. Besides, we compare our method to the state-of-the-art long-tailed learning methods and uncertainty estimation methods, respectively. For the former group, we include typical re-balancing approaches: weighted cross-entropy loss (WL) [40] and random-over-sampling (ROS) [41]. We also employ a recently proposed supervised deep clustering method (SDC) [42]. SDC first learns the class embeddings by maximizing cluster separation, and then uses a novel triplet loss to discriminate the learned embeddings. This two-stage learning protocol improves the reliability against imbalanced training data. For the latter group, we first report the performance of EDL optimized by Eq. 2, which is termed as Vanilla EDL without re-balancing the class. We also compare EDL with the other two uncertainty quantification approaches. The first approach is the Monte

Carlo Dropout method (referred to as MCDP) [22], [43], which captures model uncertainty by keeping dropout activated during testing. The other approach is deep ensemble learning (referred to as Ensemble), which quantifies uncertainty based on the outputs of multiple models [13], [44]. Although these methods have shown promise in well-curated data, they were not specifically designed for imbalanced data. To ensure a fair comparison, we implemented them using the same data augmentation techniques as EDL, namely MCDP+ROS and Ensemble+ROS. Instead of using the *DE* metric, the uncertainty measurement for non-EDL methods was the *Entropy* of the predictive probabilities [45].

For all the methods in this paper, we used a learning rate of 10^{-4}, the Adam optimizer, a batch size of 64, and a maximum epoch of 200. The best model based on the highest accuracy on the validation set was saved. ResNet-34 and DenseNet-121 were initialized with pre-trained checkpoints, while other parameters were randomly initialized. The training stops with a model saved when the best performance on the validation set is achieved. We run each experiment 5 times with different random seeds and report the average performance. All models were implemented using PyTorch 1.16, and we trained the models on a single Nvidia GPU with 64GB memory.

C. Metrics

For evaluation, we report accuracy-centric metric Rec and uncertainty-centric metrics Brier and ECE. Rec is the macro-recall on the testing set, denoted by, $Rec = \frac{1}{C}\sum_{c=1}^{C} ACC(\hat{y}^{(i)}|y^{(i)}=c)$. Brier, short for Brier Score, measures the accuracy of predicted probabilities. Specifically, the Brier Score for a sample is computed as the squared error of a predicted probability vector, $\boldsymbol{p}^{(i)}$, and the one-hot encoded true response, $\tilde{\boldsymbol{y}}^{(i)}$. That is $Brier^{(i)} = \frac{1}{C}\sum_{c=1}^{C}(p_c^{(i)} - \tilde{y}_c^{(i)})^2$. We report the average Brier Score across the whole testing set, denoted by, $Brier = \frac{1}{C}\sum_{c=1}^{C}\frac{1}{N_c}\sum_{y^{(i)}=c} Brier^{(i)}$. Rec and Brier were calculated at the macro level. We also report ECE, short for Expected Calibration Error, to measure the correspondence between predicted probabilities and empirical accuracy [46]. We partitioned the estimated confidence into $M = 10$ equal bins on the test dataset and calculated the ECE as follows: $ECE = \sum_{m}^{M}\frac{|B_m|}{N_{test}}|ACC(B_m) - conf(B_m)|$, where bin B_m covers the confidence interval $(\frac{m-1}{M}, \frac{m}{M}]$. $ACC(B_m)$ and $conf(B_m)$ are the ACC and the average predictive confidence for the samples whose predictive confidence falls within the bin B_m. Rec evaluates the overall accuracy of categorical predictions, while Brier and ECE assess the calibration of predicted probabilities.

We also evaluate two uncertainty measurement-driven applications: misclassification identification and out-of-distribution (OOD) detection [19]. We evaluate the performance by AUC_m and AUC_o for the two tasks, respectively. AUC, short for AUROC (area under the receiver operating characteristic), is used to measure the accuracy of classification. We treat the evaluation as a binary classification task: misclassified/OOD data belongs to the positive class while correctly predicted/ID data is the negative class. We conduct min-max normalization for uncertainty measurements on the testing set (for EDL methods, we use *DE*, and for other baselines, we use *Entropy*), resulting in the normalized values ranging $[0, 1]$. Those normalized uncertainty measurements are the probabilities to calculate AUC. To distinguish between near and far out-of-distribution (OOD) detection, AUC_o^n and AUC_o^f are reported, respectively.

TABLE II
PERFORMANCE COMPARISON FOR MEDICAL APPLICATIONS. THE AVERAGE RESULTS OF FIVE RUNS ARE SHOWN. THE BEST RESULTS ARE HIGHLIGHTED. THE SECOND-BEST RESULTS ARE UNDERLINED FOR COMPARISON.

	Rec↑	Brier↓	ECE↓	AUC_m↑	AUC_o^n↑	AUC_o^f↑
	Task 1: Respiratory abnormity detection					
Vanilla	0.256	0.999	0.310	0.587	0.650	0.728
WL	0.401	0.949	0.292	0.594	0.661	0.698
ROS	0.407	0.941	0.301	0.605	0.673	0.742
SDC	0.422	0.902	0.288	0.617	0.664	0.747
Vanilla EDL	0.268	0.983	0.304	0.603	0.655	0.734
EDL+WL	0.389	0.908	0.290	0.621	0.687	0.759
EDL+ROS	0.434	0.878	0.297	0.620	0.700	0.768
MCDP+ROS	0.412	0.933	0.289	0.625	0.690	0.764
Ensemble+ROS	0.431	0.929	0.286	0.628	0.699	0.769
Ours	0.422	0.797	0.163	0.640	0.727	0.785
	Task2: Skin lesion screening					
Vanilla	0.610	0.538	0.217	0.740	0.695	0.789
WL	0.689	0.457	0.159	0.784	0.665	0.891
ROS	0.727	0.441	0.110	0.801	0.693	0.927
SDC	0.730	0.439	0.112	0.813	0.705	0.927
Vanilla EDL	0.601	0.534	0.214	0.747	0.688	0.803
EDL+WL	0.678	0.511	0.153	0.798	0.694	0.882
EDL+ROS	0.735	0.428	0.105	0.830	0.701	0.896
MCDP+ROS	0.734	0.429	0.103	0.835	0.735	0.949
Ensemble+ROS	0.739	0.420	0.102	0.840	0.735	0.950
Ours	0.763	0.396	0.095	0.854	0.747	0.968
	Task 3: Heart failure prediction					
Vanilla	0.389	0.690	0.179	0.850	0.782	0.885
WL	0.715	0.480	0.073	0.608	0.690	0.766
ROS	0.717	0.482	0.071	0.597	0.681	0.758
SDC	0.732	0.476	0.073	0.600	0.692	0.770
Vanilla EDL	0.388	0.685	0.175	0.843	0.786	0.887
EDL+WL	0.585	0.521	0.123	0.622	0.788	0.893
EDL+ROS	0.690	0.478	0.062	0.848	0.790	0.920
MCDP+ROS	0.721	0.471	0.067	0.602	0.707	0.772
Ensemble+ROS	0.728	0.452	0.068	0.598	0.708	0.798
Ours	0.778	0.319	0.062	0.911	0.917	0.973

D. Results

Results are summarized in Table II and discussed below.

Task 1. The task involves a 4-class classification problem with mildly imbalanced data (refer to Table I). The first observation is that both Vanilla and Vanilla EDL struggle to perform well, while the re-balancing strategy WL and ROS significantly improve the Vanilla and Vanilla EDL across all the metrics. SDC is a strong baseline for class imbalanced data by ensuring the class margin, but it is still a deterministic model using Softmax to generate the final prediction, which indicates that the model could be over-confident for out-of-distribution data. As proven by the results, the uncertainty-aware baselines, i.e., EDL+WL, EDL+ROS, MCDP+ROS, and Ensemble+ROS, generally perform better for uncertainty-centric metrics. However,

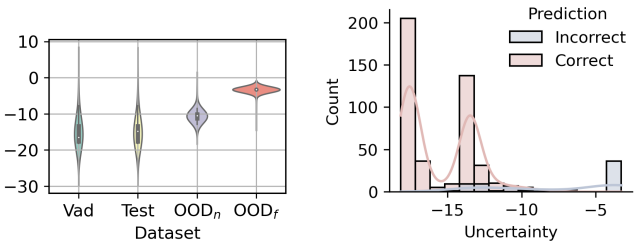
within those methods, none of them consistently outperforms the others across all metrics, highlighting the challenge of achieving accurate diagnosis accuracy and high-quality uncertainty measurements simultaneously in real-world applications. We recognize that this difficulty primarily stems from the heterogeneity of the data, as the audio recordings were collected using different stethoscopes. Thus, an effective uncertainty estimation method is necessary to accurately quantify the uncertainty from both the data and the model.

In comparison to the baselines, our class-balanced EDL approach achieves competitive results in terms of Rec. Although a Rec of 0.422 is not the best, it is very close to the best of 0.434. Yet, our method demonstrates significantly superior uncertainty measurements. Notably, we have successfully reduced ECE by 43%, indicating that our model can effectively avoid overconfident detection of respiratory abnormalities. The accuracy of detecting misclassifications and OODs is also improved by 2.4%, and 2.1% $\sim$ 3.9% compared to the second best as underlined in Table II, respectively.

Task 2. In Task 2, the training data consists of 67.1% images from healthy subjects, while the remaining data comprises six other types of lesions, exhibiting a long-tailed distribution. On this type of data, vanilla methods (Vanilla and Vanilla EDL) are vulnerable and all other methods outperform them in terms of classification and uncertainty quantification.

Among the baselines, Ensemble+ROS achieves the best performance. However, our class-balanced EDL still exhibits performance gains compared to Ensemble+ROS for all the metrics. Specifically, we can observe the improvements of 3.2% in balanced Rec, 5.7% in Brier, 6.7% in ECE, and about 2% in AUC_m, AUC_o^n, and AUC_o^f. Those suggest our classification is more accurate and our estimated uncertainty is more useful. It is also worth mentioning that the Ensemble baseline requires multiple passes during inference, making it less efficient compared to our method. These observations empirically validate the superiority of our methods over the compared baselines.

Task 3. In this task, the physiological data is highly imbalanced, with the three minority classes accounting for less than 10% of the data. From Table II, it can be observed that with such severe class imbalance, SDC achieves the highest Rec of 0.732 among the compared methods. While baselines including EDL+WL, EDL+ROS, MCDP+ROS, and Ensemble+ROS significantly improve the classification performance as measured by Rec, and reduce overconfident predictions as reflected by Brier and ECE, they fail to improve uncertainty measurement, i.e., no better AUCs. Firstly, the second best AUC_m, AUC_o^n, and AUC_o^f are 0.067, 0.848, and 0.920, showing marginal difference with 0.843, 0.786, 0.887 of vanilla EDL. Secondly, some approaches present a decline in AUCs. For example, Ensemble+ROS almost doubles the Rec but decreases the AUC_m by 29.1%. It is plausible that the baselines with weighted loss or data augmentation mechanisms can effectively reduce bias in classification, but they are unable to mitigate bias in uncertainty quantification.



(a) DE for all testing sets (Vad is the validation set). (b) DE for for in-distribution (ID) testing set.

Fig. 5. Uncertainty distribution measured by $\boldsymbol{DE}$ for heart failure prediction (Task 3).

Achieving high-quantity uncertainty estimation for this task is more challenging than the other two tasks, since Task 3 has smaller training and test data, and the OOD data shares the closest semantic information with the ID ECG data. However, even in this very challenging task, our class-balanced EDL constantly performs well on all metrics. Significantly, we increase Rec by 6.9%, decrease Brier by 29.4%, and improve the accuracy of detecting misclassifications and OODs by 5.8% $\sim$ 16.1%, respectively. All the results demonstrate that our method, which involves joint optimization of EDL posterior and prior, improves both classification performance and the quality of uncertainty for heart failure prediction simultaneously.

E. Implications of the estimated uncertainty

To gain a deeper understanding, we visualize the uncertainty distribution estimated using our method for the training, validation, and testing sets of Task 3 in Fig. 5(a). It is evident that the near and far out-of-distribution (OOD) sets exhibit larger uncertainty measurements compared to the validation and in-distribution (ID) testing sets, with the far OOD set displaying even greater uncertainty. This observation implies that an uncertainty threshold can be identified from the validation set and utilized to reject certain automatic diagnoses made by the system (as shown in Fig. 1). This approach effectively reduces the risk of misdiagnosis caused by shifts in the data distribution.

Within the ID testing set, we further divide the predictions into correct and incorrect prediction groups, and their corresponding uncertainties are displayed in Fig. 5(b). It is evident that correct predictions tend to have lower uncertainty compared to incorrect predictions. This indicates that we can effectively use uncertainty measurements to identify cases where the model cannot make automatic diagnostics and promptly refer them to clinicians for further evaluation and correction. This human-in-the-loop diagnostic pipeline can significantly enhance the overall performance of disease diagnostics while reducing the workload for clinicians compared to traditional diagnostic systems [47].

TABLE III
PERFORMANCE COMPARISON FOR THE BINARY CLASSIFICATION (A SUB-TASK OF TASK 3) UNDER DIFFERENT CLASS DISTRIBUTION.

Normal : Abnormal	Method	Rec↑	Brier↓	ECE↓	AUC_m↑	AUC_o^n↑	AUC_o^f↑
58.4% : 35.3%	Vanilla EDL	0.784	0.231	0.100	0.878	0.802	0.935
	Ours	0.803	0.219	0.009	0.891	0.844	0.937
58.4% : 17.7%	Vanilla EDL	0.617	0.346	0.178	0.679	0.712	0.835
	Ours	0.724	0.291	0.135	0.756	0.797	0.883
58.4% : 8.8%	Vanilla EDL	0.556	0.398	0.205	0.587	0.650	0.728
	Ours	0.681	0.313	0.157	0.690	0.781	0.805

F. Robustness to Different Class Imbalance Levels

To directly demonstrate the potential of our proposed method in enhancing classification and uncertainty quantification across various degrees of class imbalance, we conduct a detailed comparison using Task 3. As illustrated in Table I, the two major classes account for 58.4% (normal class) and 35.3% (R-on-T abnormal class) of the dataset. We applied random downsampling to the abnormal class and exclusively utilized data from these two classes for model training and evaluation. Table III provides a comprehensive summary of the results for the binary classification task.

The results clearly indicate a substantial degradation in performance across all metrics due to the reduced availability of the abnormal class, as the severe class imbalance significantly heightened the task's complexity. However, when compared to vanilla EDL, our method, trained using the pooling loss and the adaptive prior, exhibits greater stability in the presence of class imbalance. The relative performance gain, particularly in scenarios with more skewed class distributions, is even more remarkable. For instance, the improvement in Recall increases from 2.4% to 22.5% when the proportion of the abnormal class decreases from 35.3% to 8.8%. All of these observations collectively suggest that our method demonstrates robustness across various class imbalance levels.

G. Ablation Study

Our method consists of three key components: the novel class pooling loss calculation module $\mathcal{L}'$ (Eq. 4), the novel adaptive prior parameterized by β (Eq. 5), and the previously proposed sample-wise KL-divergence based loss function $\mathcal{L}^{(i)}$ (Eq. 2). In this section, we selectively degrade each key component to its vanilla version to understand the role of each component. Specifically, we report the performance for Ours-Dir(1), wherein our method uses the uniform prior instead of the trainable prior, Ours-Dir(η), wherein our method using the reverse class proportion as the prior instead of the trainable prior, and Ours-average, wherein our method using the average loss over a batch instead of the class pooling loss. Furthermore, we also compare our method to $\mathcal{I}$-EDL [48], a recently proposed EDL method which introduces Fisher Information Matrix to measure the informativeness of evidence carried by each sample i, according to which we can dynamically reweight the loss term $\mathcal{L}^{(i)}$ to make the model more focus on the representation learning of uncertain data. Although $\mathcal{I}$-EDL considers data uncertainty which mainly arises from the data noise and label ambiguity, it may be less effective for class imbalance, which causes model uncertainty due to data sparsity. Therefore, for the imbalanced medical data, we also implement our class pooling loss and adaptive prior mechanisms into the $\mathcal{I}$-EDL method, which we term as Ours+$\mathcal{I}$ for a comparison. The results of the above methods for Task 3 are summarized in Table IV.

TABLE IV
PERFORMANCE FOR ABLATION STUDY.

	Rec↑	Brier↓	ECE↓	AUC_m↑	AUC_o^n↑	AUC_o^f↑
Ours	0.778	0.319	0.062	0.911	0.917	0.973
Ours-Dir(η)	0.755	0.340	0.070	0.881	0.894	0.950
Ours-Dir(1)	0.694	0.401	0.079	0.876	0.878	0.927
Ours-Average	0.417	0.585	0.120	0.863	0.824	0.905
Vanilla EDL	0.388	0.685	0.175	0.843	0.786	0.887
$\mathcal{I}$-EDL	0.458	0.500	0.092	0.902	0.897	0.942
Ours+$\mathcal{I}$	0.785	0.309	0.058	0.923	0.930	0.983

From the results, it is evident that our complete method outperforms all the degraded versions, suggesting that each design in our method makes an independent contribution to the final results. Among the degraded versions, the Ours-average approach presents the most significant performance decline. This demonstrates the superiority of our proposed class pooling loss derivation in addressing the class imbalance challenge. Then, looking at the results for $\mathcal{I}$-EDL, it enhances the vanilla EDL with a notable performance improvement; however, it is outperformed by our method. This reinforces our assumption that $\mathcal{I}$-EDL can enhance learning for uncertain data but is still insufficient for the underrepresented class. When combining $\mathcal{I}$-EDL with our method, we observe a performance boost of $0.9 \sim 6.5\%$. This suggests our method can be adapted to other variants to leverage the strengths of each approach.

In summary, our method proves to be highly effective for diagnosis and uncertainty quantification in a variety of imbalanced data scenarios. It not only brings significant improvements over vanilla EDL but also outperforms many baseline methods, particularly in cases of extreme data imbalance. These results pave the way for reliable deep learning-driven health diagnosis applications in real-world settings.

VI. RELATED WORKS

A. Uncertainty Quantification

As previously discussed, Softmax-based deep classifiers are widely adopted; however, they can only quantify uncertainty from data. To enable more reliable uncertainty estimates for misclassification identification and

OOD detection, researchers have explored more advanced uncertainty-aware deep learning techniques.

Bayesian neural networks quantify the overall uncertainty by learning a distribution over the model parameters [49]. However, deriving the posterior of the model becomes intractable due to the large number of parameters in modern deep neural networks. To address this challenge, approximations such as variational inference [50], and Monte Carlo Dropout [22] have been proposed to facilitate computation. Despite the simplicity of Dropout approximation, it may not adequately capture the epistemic uncertainty arising from the model, especially when the dropout rate is low.

Deep ensembles, known as a frequentist method for uncertainty estimation, train multiple models using different subsets of the data or model initializations (Ganaie et al., 2021). While ensembles have demonstrated effectiveness, they come with increased computation and memory costs [44]. As a result, ensembles have limited applicability in real-time applications that have strict memory, time, and safety requirements [20].

In light of this, Evidential Deep Learning (EDL) has emerged as a cost-effective approach. EDL also follows the Bayesian rule, but it quantifies uncertainty by considering the distribution over predictions rather than model parameters [14]–[16], [26]. Through our experiments in Sec. V, we have also empirically demonstrated its outstanding performance in probability calibration and misclassification as well as OOD detection, in comparison with the various baselines discussed above.

B. Uncertainty for Health Applications

Due to the safety-critical nature of health applications, there has been a growing interest in incorporating uncertainty quantification for ensuring trustworthiness in the literature [2], [9], [10]. For instance, Lei et al. utilized uncertainty estimation in diagnosing diabetic retinopathy from fundus images of the eye [5], [6]. The quantified uncertainty was used for selective prediction: retaining low-uncertain outputs while referring high-uncertain predictions to doctors, involving clinicians in the loop and enhancing the system's robustness. This approach indicates that uncertainty estimates enable a human-in-the-loop medical diagnosis, helping to mitigate misdiagnosis from the model. Similarly, uncertainty-aware emotion recognition from video [51], lung disease prediction from X-rays [7], and OOD detection for skin lesion diagnostic systems [8], [9] have also been investigated. Thanks to uncertainty measurements, these models have shown superior performance and improved robustness in realistic deployment settings. Moving beyond image models, Xia et al. [10], [45], [52] have benchmarked uncertainty estimation methods on various data modalities, i.e., respiratory sounds, heart activity, brain waves, etc. Although uncertainty-aware modeling has been explored in various applications, the existing focus has mainly been on simple Softmax-based approaches that can be overly confident, or MCDP and Ensembles that require substantial computational power and are challenging to deploy in real-world scenarios. In comparison to all existing studies, we are the first to explore the state-of-the-art uncertainty quantification method EDL in health diagnostics, and our experiments have demonstrated its strong performance on there different health data modalities and applications.

C. Class Imbalance

Class imbalance is a prevalent issue in health [53] and the broader machine learning applications. A pleura of techniques have been proposed to address this problem, including three categories: information augmentation, class re-balancing, and module improvement [54]. The simplest information augmentation methods are random under-sampling (RUS) and random over-sampling (ROS) [55]. Those methods are frequently used for health data to handle the imbalance [56]–[58]. Yet, they become infeasible when the data imbalance is extreme [59]. Synthetic generation [60] or interpolation [56], [61] to increase the minority samples are also explored. However, they are sensitive to imperfections in the generated data and hard to generalize. Class re-balancing methods modify the training procedure by introducing cost-sensitive losses or scaling the classification thresholds [62]. Well-known implementations include class-balanced loss [63], focal loss [64], and recently developed contrastive loss [65] and de-biased cross-entropy loss [66]. Despite their effectiveness, those methods usually involve hyperparameters that need to be carefully tuned during training, making them difficult to generalize. Moreover, those methods are based on Softmax and thus cannot be directly adapted into the EDL framework for uncertainty quantification. Module improvement often evolves novel model designs to alleviate the bias caused by class imbalance. Ozturk et al. proposed to decouple the learning of features and the classifier: this method uses the deep clustering method to obtain features with maximum class separation and then learns the classifier by keeping the class marginal [42]. Similarly, Li et al. proposed cross-staged distilling method to prevent the classifier from being biased based on the learned features [67]. The attention mechanism was also leveraged to exploit class-agnostic global attention feature maps for the imbalanced medical data [68]. Although those methods present strong performance, they usually require more data and yield additional computational costs.

In this paper, we integrate class re-balancing method into EDL framework. Existing loss-based or threshold-based solutions are designed for the Softmax-based classifiers, which does not apply to Dirichlet-based methods like EDL. To this end, we proposed novel and easy-to-implement mechanisms that can be inherently integrated with EDL for EDL to tackle the class imbalance challenge. Extensive comparison in Sec. V-D also demonstrates the superiority of our method.

VII. Conclusion

This paper presented a systematic uncertainty quantification study to address the challenges posed by imbalanced medical data. By devising novel mechanisms for

EDL, we significantly improved its effectiveness in both classification performance and uncertainty estimation in the presence of class imbalance when applied to health diagnostics. Through extensive experiments across various data modalities and imbalance levels, the superiority of our class-balanced EDL method was demonstrated. Our study has important implications for the practical and reliable deployment of uncertainty-aware intelligent health diagnosis systems in real-world settings. Our study highlights the significance of considering class imbalance in uncertainty quantification for health diagnosis and holds crucial implications for the practical and reliable deployment of uncertainty-aware intelligent health diagnosis systems in real-world settings, providing valuable support for decision-making processes. For future work, we plan to explore the deployment of EDL for regression-based health problems with skewed training data. It is also promising to further enhance uncertainty quantification by differentiating between model uncertainty and data uncertainty under data imbalance, leading to more reliable interpretations of the outcomes.

Reference

[1] B. Kompa, J. Snoek, and A. L. Beam, "Second opinion needed: communicating uncertainty in medical machine learning," NPJ Digital Medicine, vol. 4, no. 1, pp. 1–6, 2021.

[2] U. Bhatt, J. Antorán, Y. Zhang, Q. V. Liao, P. Sattigeri, R. Fogliato, G. Melançon, R. Krishnan, J. Stanley, O. Tickoo et al., "Uncertainty as a form of transparency: Measuring, communicating, and using uncertainty," in Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society, 2021, pp. 401–413.

[3] J. Gawlikowski, C. R. N. Tassi, M. Ali, J. Lee, M. Humt, J. Feng, A. Kruspe, R. Triebel, P. Jung, R. Roscher et al., "A survey of uncertainty in deep neural networks," arXiv preprint arXiv:2107.03342, 2021.

[4] H. D. Kabir, A. Khosravi, M. A. Hosen, and S. Nahavandi, "Neural network-based uncertainty quantification: A survey of methodologies and applications," IEEE access, vol. 6, pp. 36 218–36 234, 2018.

[5] C. Leibig, V. Allken, M. S. Ayhan, P. Berens, and S. Wahl, "Leveraging uncertainty information from deep neural networks for disease detection," Scientific Reports, vol. 7, no. 1, pp. 1–14, 2017.

[6] M. Raghu, K. Blumer, R. Sayres, Z. Obermeyer, B. Kleinberg, S. Mullainathan, and J. Kleinberg, "Direct uncertainty prediction for medical second opinions," in Proceedings of the 36th International Conference on Machine Learning (ICML). PMLR, 2019, pp. 5281–5290.

[7] B. Ghoshal and A. Tucker, "Estimating uncertainty and interpretability in deep learning for coronavirus (COVID-19) detection," arXiv preprint arXiv:2003.10769, 2020.

[8] R. C. Maron, J. G. Schlager, S. Haggenmüller, C. von Kalle, J. S. Utikal, F. Meier, F. F. Gellrich, S. Hobelsberger, A. Hauschild, L. French et al., "A benchmark for neural network robustness in skin cancer classification," European Journal of Cancer, vol. 155, pp. 191–199, 2021.

[9] A. G. Pacheco, C. S. Sastry, T. Trappenberg, S. Oore, and R. A. Krohling, "On out-of-distribution detection algorithms with deep neural skin cancer classifiers," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 2020, pp. 732–733.

[10] D. Y. Kang, P. N. DeYoung, J. Tantiongloc, T. P. Coleman, and R. L. Owens, "Statistical uncertainty quantification to augment clinical decision support: a first implementation in sleep medicine," NPJ Digital Medicine, vol. 4, no. 1, p. 142, 2021.

[11] S. Afzal, M. Maqsood, F. Nazir, U. Khan, F. Aadil, K. M. Awan, I. Mehmood, and O.-Y. Song, "A data augmentation-based framework to handle class imbalance problem for alzheimer's stage detection," IEEE Access, vol. 7, pp. 115 528–115 539, 2019.

[12] P. Tschandl, C. Rosendahl, and H. Kittler, "The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions," Scientific data, vol. 5, no. 1, pp. 1–9, 2018.

[13] T. Xia, J. Han, L. Qendro, T. Dang, and C. Mascolo, "Uncertainty-aware covid-19 detection from imbalanced sound data," in Proceedings of the 22nd Annual Conference of the International Speech Communication Association (INTERSPEECH), 2021, pp. 216–220.

[14] A. Malinin and M. Gales, "Predictive uncertainty estimation via prior networks," in Proceedings of the 32nd International Conference on Neural Information Processing Systems (NeurIPS), 2018, pp. 6405–6416.

[15] M. Sensoy, L. Kaplan, and M. Kandemir, "Evidential deep learning to quantify classification uncertainty," in Proceedings of the 32nd International Conference on Neural Information Processing Systems (NeurIPS), 2018, pp. 6405–6416.

[16] B. Charpentier, D. Zügner, and S. Günnemann, "Posterior network: Uncertainty estimation without ood samples via density-based pseudo-counts," in Proceedings of the 34th International Conference on Neural Information Processing Systems (NeurIPS), 2020, pp. 1356–1367.

[17] A.-K. Kopetzki, B. Charpentier, D. Zügner, S. Giri, and S. Günnemann, "Evaluating robustness of predictive uncertainty estimation: Are dirichlet-based models reliable?" in Proceedings of the 33th International Conference on Machine Learning (ICML), 2021, pp. 5707–5718.

[18] M. Shen, Y. Bu, P. Sattigeri, S. Ghosh, S. Das, and G. Wornell, "Post-hoc uncertainty learning using a dirichlet meta-model," in Proceedings of the AAAI Conference on Artificial Intelligence (AAAI), vol. 37, no. 8, 2023, pp. 9772–9781.

[19] ——, "Post-hoc uncertainty learning using a dirichlet meta-model," arXiv preprint arXiv:2212.07359, 2022.

[20] J. Postels, M. Segu, T. Sun, L. Van Gool, F. Yu, and F. Tombari, "On the practicality of deterministic epistemic uncertainty," Proceedings of the 36th International Conference on Neural Information Processing Systems (NeurIPS), vol. 35, 2022.

[21] D. Hendrycks and K. Gimpel, "A baseline for detecting misclassified and out-of-distribution examples in neural networks," in International Conference on Learning Representations (ICLR), 2016.

[22] Y. Gal and Z. Ghahramani, "Dropout as a bayesian approximation: Representing model uncertainty in deep learning," in Proceedings of the 33th International Conference on Machine Learning (ICML), 2016, pp. 1050–1059.

[23] R. Soni, N. Shah, and J. D. Moore, "Fine-grained uncertainty modeling in neural networks," arXiv preprint arXiv:2002.04205, 2020.

[24] T. Hastie, R. Tibshirani, J. H. Friedman, and J. H. Friedman, The elements of statistical learning: data mining, inference, and prediction. Springer, 2009, vol. 2.

[25] K. P. Murphy, Machine learning: a probabilistic perspective. MIT press, 2012.

[26] D. Ulmer, "A survey on evidential deep learning for single-pass uncertainty estimation," arXiv preprint arXiv:2110.03051, 2021.

[27] S. Tammina, "Transfer learning using vgg-16 with deep convolutional neural network for classifying images," International Journal of Scientific and Research Publications (IJSRP), vol. 9, no. 10, pp. 143–150, 2019.

[28] Q. Zou, S. Xie, Z. Lin, M. Wu, and Y. Ju, "Finding the best classification threshold in imbalanced classification," Big Data Research, vol. 5, pp. 2–8, 2016.

[29] S. Wang, L. L. Minku, N. Chawla, and X. Yao, "Learning from data streams and class imbalance," Connection Science, vol. 31, no. 2, pp. 103–104, 2019.

[30] V. Bengs, E. Hüllermeier, and W. Waegeman, "Pitfalls of epistemic uncertainty quantification through loss minimisation," in Proceedings of the 36th International Conference on Neural Information Processing Systems (NeurIPS), 2022.

[31] S. Gairola, F. Tom, N. Kwatra, and M. Jain, "Respirenet: A deep neural network for accurately detecting abnormal lung sounds in limited data setting," in Proceedings of the 43rd Annual International Conference of the IEEE Engineering in Medicine

& Biology Society (EMBC), 2021, pp. 527–530.
[32] B. M. Rocha, D. Filos, L. Mendes, G. Serbes, S. Ulukaya, Y. P. Kahya, N. Jakovljevic, T. L. Turukalo, I. M. Vogiatzis, E. Perantoni et al., "An open access database for the evaluation of respiratory sound classification algorithms," Physiological measurement, vol. 40, no. 3, p. 035001, 2019.
[33] M. Fraiwan, L. Fraiwan, B. Khassawneh, and A. Ibnian, "A dataset of lung sounds recorded from the chest wall using an electronic stethoscope," Data in Brief, vol. 35, p. 106913, 2021.
[34] T. Iqbal, "ARCA23K," 2021. [Online]. Available: https://zenodo.org/record/5117901#.YkCsRk3MJPY
[35] N. C. Codella, D. Gutman, M. E. Celebi, B. Helba, M. A. Marchetti, S. W. Dusza, A. Kalloo, K. Liopyris, N. Mishra, H. Kittler et al., "Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (isbi)," in Proceedings of the 2018 IEEE 15th international symposium on biomedical imaging (ISBI), 2018, pp. 168–172.
[36] R. Avanzato and F. Beritelli, "Automatic ecg diagnosis using convolutional neural network," Electronics, vol. 9, no. 6, p. 951, 2020.
[37] C. Y, "ECG50000." [Online]. Available: https://timeseriesclassification.com/description.php?Dataset=ECG5000
[38] R. T. Olszewski, Generalized feature extraction for structural pattern recognition in time-series data. Carnegie Mellon University, 2001.
[39] K. E, "NonInvasiveFetalECGThorax1." [Online]. Available: https://timeseriesclassification.com/description.php?Dataset=NonInvasiveFetalECGThorax1
[40] Y. S. Aurelio, G. M. de Almeida, C. L. de Castro, and A. P. Braga, "Learning from imbalanced data sets with weighted cross-entropy function," Neural processing letters, vol. 50, no. 2, pp. 1937–1949, 2019.
[41] M. S. Shelke, P. R. Deshmukh, and V. K. Shandilya, "A review on imbalanced data handling using undersampling and oversampling technique," Int. J. Recent Trends Eng. Res, vol. 3, no. 4, pp. 444–449, 2017.
[42] Ş. Öztürk and T. Çukur, "Deep clustering via center-oriented margin free-triplet loss for skin lesion detection in highly imbalanced datasets," IEEE Journal of Biomedical and Health Informatics, vol. 26, no. 9, pp. 4679–4690, 2022.
[43] A. Lemay, K. Hoebel, C. P. Bridge, B. Befano, S. De Sanjosé, D. Egemen, A. C. Rodriguez, M. Schiffman, J. P. Campbell, and J. Kalpathy-Cramer, "Improving the repeatability of deep learning models with monte carlo dropout," npj Digital Medicine, vol. 5, no. 1, pp. 1–11, 2022.
[44] B. Lakshminarayanan, A. Pritzel, and C. Blundell, "Simple and scalable predictive uncertainty estimation using deep ensembles," in Proceedings of the 31st International Conference on Neural Information Processing Systems (NeurIPS), 2017, pp. 6405–6416.
[45] L. Qendro, A. Campbell, P. Lio, and C. Mascolo, "Early exit ensembles for uncertainty quantification," in Proceedings of the Machine Learning for Health (ML4H), 2021, pp. 181–195.
[46] Y. Ovadia, E. Fertig, J. Ren, Z. Nado, D. Sculley, S. Nowozin, J. Dillon, B. Lakshminarayanan, and J. Snoek, "Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift," 2019, pp. 1–13.
[47] K. Dvijotham, J. Winkens, M. Barsbey, S. Ghaisas, N. Pawlowski, R. Stanforth, P. MacWilliams, Z. Ahmed, S. Azizi, Y. Bachrach et al., "Enhancing the reliability and accuracy of ai-enabled diagnosis via complementarity-driven deferral to clinicians (codoc)," Nature Medicine, vol. 29, no. 1, p. 1814–1820, 2023.
[48] D. Danruo, G. Chen, Y. Yang, F. Liu, and P.-A. Heng, "Uncertainty estimation by fisher information-based evidential deep learning," pp. 1050–1059, 2023.
[49] C. Blundell, J. Cornebise, K. Kavukcuoglu, and D. Wierstra, "Weight uncertainty in neural network," in Proceedings of the 32nd International Conference on Machine Learning (ICML), 2015, pp. 1613–1622.
[50] D. M. Blei, A. Kucukelbir, and J. D. McAuliffe, "Variational inference: A review for statisticians," Journal of the American statistical Association, vol. 112, no. 518, pp. 859–877, 2017.
[51] J. Han, Z. Zhang, M. Schmitt, M. Pantic, and B. Schuller, "From hard to soft: Towards more human-like emotion recognition by modelling the perception uncertainty," in Proceedings of the 25th ACM International Conference on Multimedia (ACM MM), 2017, pp. 890–897.
[52] T. Xia, J. Han, and C. Mascolo, "Benchmarking uncertainty qualification on biosignal classification tasks under dataset shift," arXiv preprint arXiv:2112.09196, 2021.
[53] M. A. Mazurowski, P. A. Habas, J. M. Zurada, J. Y. Lo, J. A. Baker, and G. D. Tourassi, "Training neural network classifiers for medical decision making: The effects of imbalanced datasets on classification performance," Neural networks, vol. 21, no. 2-3, pp. 427–436, 2008.
[54] Y. Zhang, B. Kang, B. Hooi, S. Yan, and J. Feng, "Deep long-tailed learning: A survey," IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023.
[55] J. Van Hulse, T. M. Khoshgoftaar, and A. Napolitano, "Experimental perspectives on learning from imbalanced data," in Proceedings of the 24th International Conference on Machine Learning (ICML), 2007, pp. 935–942.
[56] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: synthetic minority over-sampling technique," Journal of artificial intelligence research, vol. 16, pp. 321–357, 2002.
[57] M. M. Rahman and D. N. Davis, "Addressing the class imbalance problem in medical datasets," International Journal of Machine Learning and Computing, vol. 3, no. 2, p. 224, 2013.
[58] J. Han, C. Brown, J. Chauhan, A. Grammenos, A. Hasthanasombat, D. Spathis, T. Xia, P. Cicuta, and C. Mascolo, "Exploring automatic covid-19 diagnosis via voice and symptoms from crowdsourced data," in Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2021, pp. 8328–8332.
[59] J. M. Johnson and T. M. Khoshgoftaar, "Survey on deep learning with class imbalance," Journal of Big Data, vol. 6, no. 1, pp. 1–54, 2019.
[60] H. He, Y. Bai, E. A. Garcia, and S. Li, "Adasyn: Adaptive synthetic sampling approach for imbalanced learning," in 2008 IEEE international joint conference on neural networks (IEEE world congress on computational intelligence). IEEE, 2008, pp. 1322–1328.
[61] A. Galdran, G. Carneiro, and M. A. González Ballester, "Balanced-mixup for highly imbalanced medical image classification," in Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part V 24. Springer, 2021, pp. 323–333.
[62] S. Rajaraman, P. Ganesan, and S. Antani, "Deep learning model calibration for improving performance in class-imbalanced medical image classification tasks," PloS one, vol. 17, no. 1, p. e0262838, 2022.
[63] Y. Cui, M. Jia, T.-Y. Lin, Y. Song, and S. Belongie, "Class-balanced loss based on effective number of samples," in Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR), 2019, pp. 9268–9277.
[64] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in Proceedings of the IEEE international conference on computer vision, 2017, pp. 2980–2988.
[65] Z. Yang, J. Pan, Y. Yang, X. Shi, H.-Y. Zhou, Z. Zhang, and C. Bian, "Proco: Prototype-aware contrastive learning for long-tailed medical image classification," in International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, 2022, pp. 173–182.
[66] L. Luo, D. Xu, H. Chen, T.-T. Wong, and P.-A. Heng, "Pseudo bias-balanced learning for debiased chest x-ray classification," in International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, 2022, pp. 621–631.
[67] J. Li, G. Chen, H. Mao, D. Deng, D. Li, J. Hao, Q. Dou, and P.-A. Heng, "Flat-aware cross-stage distilled framework for imbalanced medical image classification," in International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, 2022, pp. 217–226.
[68] A. He, T. Li, N. Li, K. Wang, and H. Fu, "Cabnet: Category attention block for imbalanced diabetic retinopathy grading," IEEE Transactions on Medical Imaging, vol. 40, no. 1, pp. 143–153, 2020.