Human Uncertainty Inference via Deterministic Ensemble Neural Networks

Yujin Cha[1], **Sang Wan Lee**[1,2,3]
[1]Department of Bio and Brain Engineering, Korea Advanced Institute of Science and Technology (KAIST)
[2]Program of Brain and Cognitive Engineering, Korea Advanced Institute of Science and Technology (KAIST)
[3]Center for Neuroscience-inspired Artificial Intelligence, Korea Advanced Institute of Science and Technology (KAIST)
{chayj, sangwan}@kaist.ac.kr

Abstract

The estimation and inference of human predictive uncertainty have great potential to improve the sampling efficiency and prediction reliability of human-in-the-loop systems for smart healthcare, smart education, and human-computer interactions. Predictive uncertainty in humans is highly interpretable, but its measurement is poorly accessible. Contrarily, the predictive uncertainty of machine learning models, albeit with poor interpretability, is relatively easily accessible. Here, we demonstrate that the poor accessibility of human uncertainty can be resolved by exploiting simple and universally accessible deterministic neural networks. We propose a new model for human uncertainty inference, called proxy ensemble network (PEN). Simulations with a few benchmark datasets demonstrated that the model can efficiently learn human uncertainty from a small amount of data. To show its applicability in real-world problems, we performed behavioral experiments, in which 64 physicians classified medical images and reported their level of confidence. We showed that the PEN could predict both the uncertainty range and diagnoses given by subjects with high accuracy. Our results demonstrate the ability of machine learning in guiding human decision making; it can also help humans in learning more efficiently and accurately. To the best of our knowledge, this is the first study that explored the possibility of accessing human uncertainty via the lens of deterministic neural networks.

1 Introduction

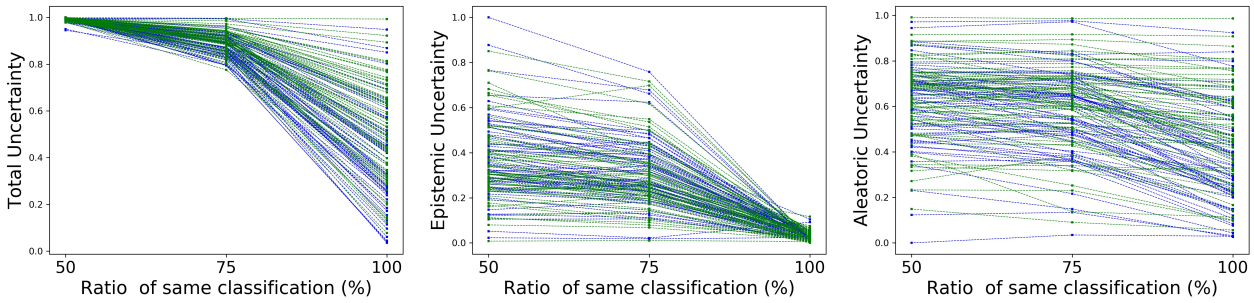
Sample efficient learning involves the ability to distinguish between "what it knows and what it does not." This can be quantified in the form of an uncertainty. An uncertainty is caused by a limited amount of available data or the limited ability of the learning agent to glean information from the given data. By assessing an uncertainty during decision making, an agent can choose whether to decide by itself or "go to oracle."(Cohn, Ghahramani, and Jordan 1996). Furthermore, the evaluation of uncertainty enables the exploration of learning strategies for improving sample efficiency under limited resource conditions.

Uncertainty in information is dealt with by various types of learning such as active and one-shot learning (Gal, Islam, and Ghahramani 2017; Lee, O'Doherty, and Shimojo 2015; Tong 2001; Peterson et al. 2019). Also, it is useful for human and probabilistic models in which different choices can be made from the same input data. A challenge in these applications is that sometimes an agent provides an accurate response by coincidence or, in another situation, it does not provide reliable predictions even if it has learned adequately. It is, therefore, important to consider the uncertainty as well as accuracy in the performance assessment of learning agents.

Although the information on uncertainty is available with respect to humans and machine learning (ML) models, there is a sharp contrast between them. First, quantifying the amount of uncertainty in "deep" heavy ML models is a challenging task. The Bayesian neural network (BNN), which uses the prior distribution of model parameters to calculate their posterior distribution for the given data, can efficiently estimate the uncertainty of deep neural networks (DNNs); however, it is not highly scalable (Gal 2016). For non-Bayesian ML models, the Monte-Carlo (MC) dropout and ensemble methods could be pragmatic solutions (Gal and Ghahramani 2016; Lakshminarayanan, Pritzel, and Blundell 2017; Liu et al. 2019). Thus, the amount of uncertainty can be assessed at a low cost. In other words, ML models are easily accessible despite poor interpretability.

Obtaining the uncertainty information in human decision-making is costly and labor-intensive. Also, other variables like the average level of uncertainty may affect uncertainty indicating that it is poorly accessible; however it is not difficult to interpret in comparison to the ML model (Barthelmé and Mamassian 2009), posing a significant challenge in the optimization of human-in-the-loop systems for smart healthcare, smart education, and human-computer interactions.

To resolve the poor accessibility of human uncertainty, we exploit the high accessibility of deterministic ML algorithms. We propose a "proxy ensemble network" (PEN), a neural network model that performs inference on human uncertainty from a small amount of data. We perform simulations to show the efficiency of learning and demonstrate the applicability of the algorithm to real-world problems through behavioral experiments. Note that while previous studies have attempted to compare the information processing of artificial neural networks (ANNs) with human cortical information processing for image recognition, our study shows that deterministic neural networks can accurately approximate the uncertainty of human experts such as physi-



(a) Total Uncertainty (b) Epistemic Uncertainty (c) Aleatoric Uncertainty

Figure 1: In our behavioral experiment, 64 physicians were presented with 50 test images 4 times in different trials. They were instructed to submit their binary classification results and confidence level. (See Section. 4.4) Uncertainty was calculated for each image and each subject (See Section 3.2). The 50 images were grouped according to the consistency of the response (100%, 75%, and 50%). Each line shows the average uncertainty for each subject (Blue line: CXR-A, Green line: CXR-B).

cians. The contributions of our study can be summarized as follows.

- We examined human uncertainty from a BNN perspective and showed that decision uncertainty can be inferred by a deterministic ensemble of ANNs. This suggests that the ability of the ANN to assess uncertainty can compensate for the poor accessibility of human uncertainty.
- Our results show that ML can not only guide human perceptual decision making, but it can also help humans learn more efficiently and accurately.
- We show the applicability of PENs in an area of smart healthcare. We performed a large-scale behavioral experiment in which 64 physicians (M.D.) performed[1] medical image classification, including difficult cases to make an accurate diagnosis, and reported the uncertainty level of each decision.

2 Related Works

2.1 Uncertainty in Deep Neural Networks

Designing an algorithm for calculating the uncertainty of DNNs is challenging. The total amount of uncertainty is defined as the sum of the predictive, model (epistemic), and data (aleatoric) uncertainties (Liu et al. 2019; Der Kiureghian and Ditlevsen 2009; Kingma, Salimans, and Welling 2015; Kendall and Gal 2017; Malinin and Gales 2018; Chai 2018). First, the model uncertainty is caused by the inaccurate estimation of model parameters; hence, it can be resolved with training. The model uncertainty is estimated using the posterior distribution of the model parameters of a given dataset. Recent studies used the Bayesian inference (BI) or MC sampling. The BI method can directly calculate the uncertainty of a model in principle, but its applicability is limited owing to poor scalability. Alternatively, the MC sampling method can approximate the posterior distribution using an ensemble network or a MC dropout strategy during test time (Gal and Ghahramani 2016; Kingma, Salimans, and Welling 2015). Second, data uncertainty arises from data bias or sensory noise. To quantify data uncertainty, a method to compute the variance of data by adding the maximum-likelihood loss during training was proposed (Kendall and Gal 2017). Another approach is a test time augmentation method (Wang et al. 2018).

[1]The dataset and details of experiments can be downloaded from : https://github.com/brain-machine-intelligence/PEN

2.2 Computational Models for Human Perception

Recent studies have compared the information processing of ANNs with human sensory processing. For example, the relationship between human subjective uncertainty and objective uncertainty in the perception of visual stimuli was examined (Barthelmé and Mamassian 2009). Several studies demonstrated the importance of stochasticity in DNNs for developing computer vision systems and computational models of human perception (McClure 2018; Nakada, Chen, and Terzopoulos 2018). A few studies examined the optimization of the structure of DNNs to better understand the computational mechanisms of human visual perception(Yamins et al. 2013, 2014). Some studies explored the concept of computational constructs of human uncertainty (Grinband, Hirsch, and Ferrera 2006; Hramov et al. 2018; van Bergen and Jehee 2019); however, little is known about human uncertainty inference.

3 Proxy Ensemble Network (PEN)

Here, we present a new model, called PEN, that learns to predict the output and uncertainty of a BNN model for new data in the test phase. First, we formulate our research question (Section 3.1). As prerequisites for the implementation, we examine the distinctive characteristics of uncertainty (Section 3.2). We then discuss the possibility of quantifying the uncertainty of the BNN based on a function of models and data (Section 3.3), by inferring the expectation and range of the model outputs (Section 3.4). We then provide a justification for why and how the BNN can model the uncertain nature of human behavior (Section 3.5). Finally, the detailed PEN process is summarized in Sections 3.6.

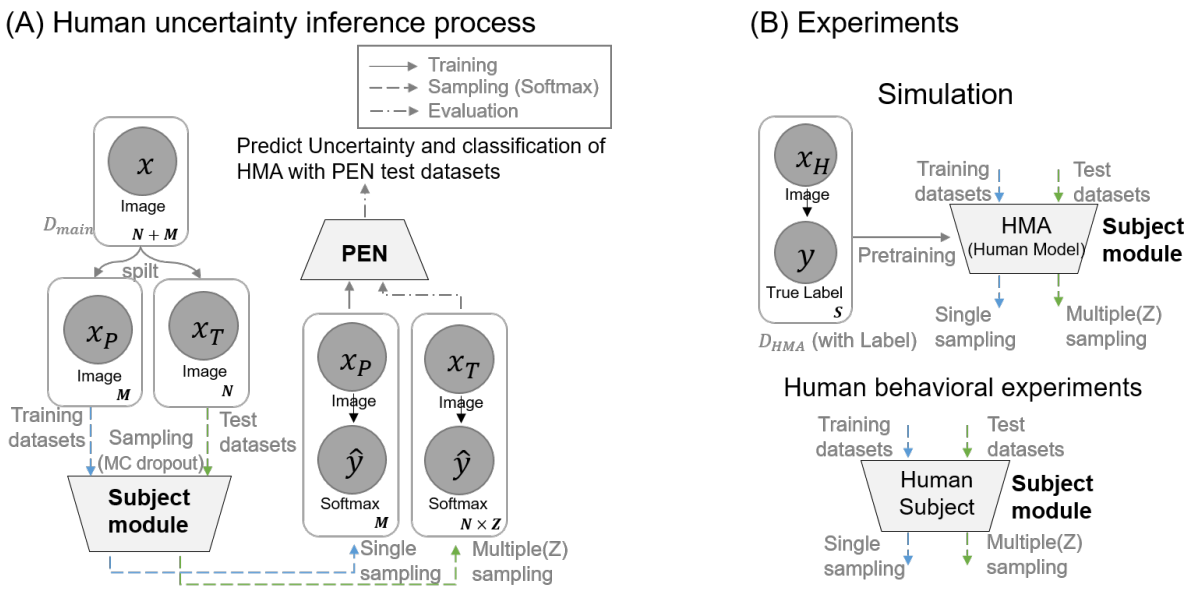


Figure 2: Overview of the proposed system : (A) Proposed framework of human uncertainty inference. (B) Simulation or human behavioral experiment

3.1 Problem Formulation

Suppose that there is a Bayesian CNN model that is trained arbitrarily for multi-class classification (K classes), referred to as an "original" model here. We show a simple process of finding a "proxy" model that enables us to predict the classification output of an original model and the range of uncertainty for untrained data. We consider i.i.d input datasets $\mathbf{X} = \{\mathbf{x}_n\}_{n=1}^{N}$. The softmax (Luce 1959) output of the original model for image $\mathbf{x}$ can be expressed as $\hat{\mathbf{y}}$, which is a K-dimensional probability vector. This provides the model output set, $\widehat{\mathbf{D}} = \{\mathbf{x}_n, \hat{\mathbf{y}}_n\}_{n=1}^{N}$. Because the output of the BNN depends on the neural weights sampled from the weight distribution, the model can provide a different $\widehat{\mathbf{D}}$ even for the same input set $\mathbf{X}$(Chai 2018). We use $\mathbf{w}$ to denote the weights of the original model that are sampled differently each time. Note that $\mathbf{w}$ can be considered as a random variable defined in the model weight probability space, $\mathcal{W}$. The probability of $\mathbf{w}$ for the given $\mathcal{W}$ is expressed as $\mathrm{p}(\mathbf{w}|\mathcal{W})$. We refer to the model as a function, $\mathbf{f}$, and use $\hat{\mathbf{y}} = \mathbf{f}^{\mathbf{w}}(\mathbf{x})$ to indicate the dependence of $\mathbf{f}$ on $\mathbf{x}$ and $\mathbf{w}$. We then train a proxy model to maximize the likelihood of the probabilistic distribution $\mathrm{p}(\hat{\mathbf{y}}|\mathbf{x};\theta)$ with a subset of $\widehat{\mathbf{D}}$, where θ is the deterministic parameter of the proxy model, by finding the maximizer, θ^*, that satisfies $argmax_\theta p(\tilde{\mathbf{y}}|\mathbf{x};\theta)$, and use $\tilde{\mathbf{y}} = \mathbf{f}^{\theta}(\mathbf{x})$ to indicate the dependence of $\mathbf{f}$ on $\mathbf{x}$ and θ. Following the Bayesian setting in which a collection of networks forms an ensemble $\theta \in \Theta$, the predicted distribution of the new output, $\widehat{\mathbf{y}'}$, for the new input $\mathbf{x}'$ is given by

$$p(\tilde{\mathbf{y}'}|\mathbf{x}';\mathbf{\Theta}) \approx p(\tilde{\mathbf{y}'}|\mathbf{x}';\widehat{\mathbf{D}}) = \int_{\Omega} p(\tilde{\mathbf{y}'}|\mathbf{x}';\theta)p(\theta|\widehat{\mathbf{D}})d\theta. \tag{1}$$

We argue that both the expectation of $\mathbf{f}^{\mathbf{w}}(\mathbf{x}')$ and the range of uncertainty, which is expressed as entropy $\mathbb{H}[\widehat{\mathbf{y}'}|\mathbf{x}';\mathbf{w}]$, can be predicted by estimating the probability distribution of empirical $p(\widehat{\mathbf{y}'}|\mathbf{x}';\mathbf{\Theta})$; this distribution can be modeled as an ensemble of deterministic neural networks composed of the sub-optimally trained parameter θ. We call this ensemble a PEN. Note that a PEN is not meant for learning the correct (ground truth) label for an input; it is intended to provide the uncertainty approximation of the original model by learning the probability distribution of the output labels given by the BNN model or a human subject (Section 3.5).

3.2 Uncertainty Estimation for BNNs

The total uncertainty of a BNN can be estimated by repeated sampling, and this can be decomposed into aleatoric and epistemic uncertainty.

Total Uncertainty. In some studies, uncertainty is measured by the entropy of the softmax function output (Gal, Islam, and Ghahramani 2017). The entropy can be obtained with one sampling for input $\mathbf{x}$ in the BNN as follows.

$$\mathbb{H}[\hat{\mathbf{y}}|\mathbf{x};\mathbf{w}] = -\sum_{k} p(\hat{\mathbf{y}} = k|\mathbf{x};\mathbf{w}) \log p(\hat{\mathbf{y}} = k|\mathbf{x};\mathbf{w}), \tag{2}$$

where $p(\hat{\mathbf{y}} = k|\mathbf{x};\mathbf{w})$ is the predicted softmax output for class k using weights $\mathbf{w}$ from $p(\mathbf{w})$ for input $\mathbf{x}$. Similarly, the total (predictive) uncertainty for input $\mathbf{x}$ in the model is given by the following information entropy (Chai 2018):

$$\mathbb{H}[\hat{\mathbf{y}}|\mathbf{x};\mathcal{W}] = -\sum_{k} p(\hat{\mathbf{y}} = k|\mathbf{x};\mathcal{W}) \log p(\hat{\mathbf{y}} = k|\mathbf{x};\mathcal{W}). \tag{3}$$

However, in practice it is impossible to integrate over $\mathbf{w}$ to estimate $\mathcal{W}$. Alternatively, we can calculate the amount of

uncertainty by sampling from a variational distribution of weights. For instance, M repetitive samplings can be performed to estimate the total amount of uncertainty.

$$\mathbb{H}[\hat{\mathbf{y}}|\mathbf{x};\mathcal{W}] \approx -\sum_k \Big(\frac{1}{M}\sum_m p(\hat{\mathbf{y}}=k|\mathbf{x};\mathbf{w}^{(\mathbf{m})})\Big) \log\Big(\frac{1}{M}\sum_m p(\hat{\mathbf{y}}=k|\mathbf{x};\mathbf{w}^{(\mathbf{m})})\Big), \quad (4)$$

where $\mathbf{w}^{(\mathbf{m})}$ is the m-th sample of weights.

Aleatoric Uncertainty. A deterministic neural network does not offer any information about the output distribution for the training data, thereby making it difficult to collect the confidence information (Gal and Ghahramani 2016). This type of uncertainty is called the aleatoric uncertainty, which is caused by the uncertain nature of data or sensory noise (Hüllermeier and Waegeman 2019). Because it is impossible to accurately measure the aleatoric uncertainty $\mathbb{E}_{\mathbf{w}\sim p(\mathbf{w}|\mathcal{W})}\big[\mathbb{H}[\hat{\mathbf{y}}|\mathbf{x};\mathbf{w}]\big]$, we use repetitive sampling again to estimate the average entropy across different weights, $\mathbf{w}^{(\mathbf{m})}$, associated with each sampling.

$$\mathbb{E}_{\mathbf{w}\sim p(\mathbf{w}|\mathcal{W})}\big[\mathbb{H}[\hat{\mathbf{y}}|\mathbf{x};\mathbf{w}]\big] \approx -\frac{1}{M}\sum_m\sum_k p(\hat{\mathbf{y}}=k|\mathbf{x};\mathbf{w}^{(\mathbf{m})})\log p(\hat{\mathbf{y}}=k|\mathbf{x};\mathbf{w}^{(\mathbf{m})}). \quad (5)$$

Epistemic Uncertainty. In the BNN, epistemic uncertainty refers to the type of uncertainty in the model parameters. It is given by the difference between the total uncertainty and aleatoric uncertainty:

$$\mathbb{I}[\hat{\mathbf{y}},\mathbf{w}|\mathbf{x};\mathcal{W}] = \mathbb{H}[\hat{\mathbf{y}}|\mathbf{x};\mathcal{W}] - \mathbb{E}_{\mathbf{w}\sim p(\mathbf{w}|\mathcal{W})}\big[\mathbb{H}[\hat{\mathbf{y}}|\mathbf{x};\mathbf{w}]\big]. \quad (6)$$

3.3 Estimating Uncertainty with Function of Models and Data

Uncertainty can be expressed as a function of the model ($\mathcal{W}$) and data ($\mathbf{x}$). Following the discussion in Section 3.2, the estimate of uncertainty converges to a constant as the number of sampling, M, increases. This implies that when M$\rightarrow\infty$, the uncertainty terms become a deterministic function. In such cases, the total uncertainty, $\mathbf{U_T}(\mathbf{x};\mathcal{W})$, aleatoric uncertainty, $\mathbf{U_A}(\mathbf{x};\mathcal{W})$, and epistemic uncertainty, $\mathbf{U_E}(\mathbf{x};\mathcal{W})$ are defined as follows.

$$\mathbf{U_T}(\mathbf{x};\mathcal{W}) \equiv \mathbb{H}[\hat{\mathbf{y}}|\mathbf{x};\mathcal{W}], \quad (7)$$

$$\mathbf{U_A}(\mathbf{x};\mathcal{W}) \equiv \mathbb{E}_{\mathbf{w}\sim p(\mathbf{w}|\mathcal{W})}\big[\mathbb{H}[\hat{\mathbf{y}}|\mathbf{x};\mathbf{w}]\big], \quad (8)$$

$$\mathbf{U_E}(\mathbf{x};\mathcal{W}) \equiv \mathbb{I}[\hat{\mathbf{y}},\mathbf{w}|\mathbf{x};\mathcal{W}]. \quad (9)$$

Similarly, each uncertainty estimated by a finite number of sampling can be defined as $\widehat{\mathbf{U_T}}(\mathbf{x};\mathcal{W})$, $\widehat{\mathbf{U_A}}(\mathbf{x};\mathcal{W})$ and $\widehat{\mathbf{U_E}}(\mathbf{x};\mathcal{W})$, respectively. Similarly, the mean vector $\mathbf{f}^{\mathcal{W}}(\mathbf{x})$ can be obtained by sampling $\mathbf{f}^{\mathbf{w}}(\mathbf{x})$ infinitely.

The entropy calculated from $\mathbf{f}^{\mathcal{W}}(\mathbf{x})$ is equal to $\mathbf{U_T}(\mathbf{x};\mathcal{W})$. If $\mathbf{f}^{\mathbf{w}}(\mathbf{x})$ is sampled only once, instead of M times, the output $\hat{\mathbf{y}}$ and its entropy are limited within a certain range. The theoretical minimum value of the entropy of $\hat{\mathbf{y}}$ is equal to $\mathbf{U_A}(\mathbf{x};\mathcal{W})$, assuming that there is no uncertainty in the model parameters. Moreover, assuming $p(\mathbf{w}|\mathcal{W})$ follows a Gaussian distribution such that $\mathbf{f}^{\mathbf{w}}(\mathbf{x})$ will have the highest likelihood value at $\mathbf{f}^{\mathcal{W}}(\mathbf{x})$; thus, the expectation of the entropy is equal to $\mathbf{U_T}(\mathbf{x};\mathcal{W})$. Because $\mathbf{U_T}(\mathbf{x};\mathcal{W}) - \mathbf{U_E}(\mathbf{x};\mathcal{W}) = \mathbf{U_A}(\mathbf{x};\mathcal{W})$, the entropy range is [$\mathbf{U_T}(\mathbf{x};\mathcal{W})$-$\mathbf{U_E}(\mathbf{x};\mathcal{W})$, $\mathbf{U_T}(\mathbf{x};\mathcal{W})$ + $\mathbf{U_E}(\mathbf{x};\mathcal{W})$], and $\hat{\mathbf{y}}$ is an output vector that satisfies this entropy condition.

In other words, the uncertainty can be defined as the entropy of the mean vector of $\hat{\mathbf{y}}$ obtained by infinite sampling. When obtained with finite sampling, it corresponds to an estimated uncertainty. The output boundary of the random variable corresponds to the range of entropy value obtained with single sampling.

3.4 PEN: Inferring the Expectation and Range of Original Model Outputs

The classification for $\mathbf{x}'$ of the original model can be predicted with $\mathbf{f}^{\mathbf{\Theta}}(\mathbf{x}')$, which is the simple average of vectors (denoted as $\tilde{\mathbf{y}'}$) outputs by the ensemble members for $\mathbf{x}'$. To predict the entropy range of the original model, the uncertainty of the PEN should be considered. The total uncertainty can be calculated at the ensemble level for unseen data $\mathbf{x}'$ as follows.

$$\mathbb{H}[\tilde{\mathbf{y}'}|\mathbf{x}';\mathbf{\Theta}] = \sum_k \Big(\frac{1}{M}\sum_m p(\tilde{\mathbf{y}'}=k|\mathbf{x}';\theta^{(\mathbf{m})})\Big)\times \log\Big(\frac{1}{M}\sum_m p(\tilde{\mathbf{y}'}=k|\mathbf{x}';\theta^{(\mathbf{m})})\Big), \quad (10)$$

where m is the ensemble number, and M is the total number of members. The total uncertainty of the ensemble for $\mathbf{x}'$ can be decomposed into the uncertainty in the original model, the epistemic uncertainty of the ensemble, and the aleatoric uncertainty of the ensemble as follows.

$$\mathbb{H}[\tilde{\mathbf{y}'}|\mathbf{x}';\mathbf{\Theta}] \approx \mathbb{H}[\hat{\mathbf{y}'}|\mathbf{x}';\mathcal{W}] + \mathbf{U_E}(\mathbf{x}';\Theta) + \mathbf{U_A}(\mathbf{x}';\Theta). \quad (11)$$

The likelihood, $\mathbb{L}(\mathbf{x}|\hat{\mathbf{y}})$, of $\widehat{\mathbf{D}}$'s constituent tuples $(\mathbf{x},\hat{\mathbf{y}})$ sampled from the original model can have various distributions. When sampling $\mathbf{f}^{\mathbf{w}}(\mathbf{x})$, if $\mathbf{U_E}(\mathbf{x};\mathcal{W})$ for $\mathbf{x}$ is small, it is likely to be sampled at a point close to $\mathbf{f}^{\mathcal{W}}(\mathbf{x})$ with high likelihood, as discussed in the section 3.3. However, if $\mathbf{U_E}(\mathbf{x};\mathcal{W})$ is large, it is more likely to be sampled at a point farther from $\mathbf{f}^{\mathcal{W}}(\mathbf{x})$. Thus, it is highly likely that the likelihood of the sampled $\mathbf{f}^{\mathbf{w}}(\mathbf{x})$ is inversely proportional to that of $\mathbf{U_E}(\mathbf{x};\mathcal{W})$. At the ensemble level of a PEN, all data for $\widehat{\mathbf{D}}$ are included in the learning stage, but the ensemble member networks that constitute the PEN learn different subsets of $\widehat{\mathbf{D}}$. Suppose that a member of the ensemble (θ') learns $(\mathbf{x},\hat{\mathbf{y}})$, which is sampled with low likelihood and predicts $\mathbf{x}^*$ close to $\mathbf{x}$ in terms of data distance. The output $\mathbf{f}^{\theta'}(\mathbf{x}^*)$ of the PEN member (θ') is close to $\hat{\mathbf{y}}$, but the member (θ''), not participating in the learning of $(\mathbf{x},\hat{\mathbf{y}})$, outputs $\mathbf{f}^{\theta''}(\mathbf{x}^*)$ close to $\mathbf{f}^{\mathcal{W}}(\mathbf{x}^*)$. This is because member θ'' only learns data excluding $\mathbf{x}$ and makes predictions about $\mathbf{x}^*$; therefore, it will output a prediction, $\mathbf{f}^{\mathcal{W}}(\mathbf{x}^*)$, with a high likelihood instead of a result that approaches $\tilde{\mathbf{y}'}$, which is the expected output of the original model. The difference in outputs of

the PEN members is reflected in $\mathbf{U_E}(\mathbf{x}^*;\Theta)$. These observations motivated us to use this difference for quantifying the entropy range, $\mathbf{U_E}(\mathbf{x}^*;\Theta)$, of the original model. Because the ground truth label can be seen as an output with zero entropy, the aleatoric uncertainty in the original model can be viewed as the difference between the entropy of the model and the ground truth label of the data (herein referred to as *absolute data uncertainty*). From the perspective of the PEN, the ground truth label is not a label with zero entropy, but a label reflecting the inherent data uncertainty perceived by the original model. The aleatoric uncertainty of the PEN denotes the absolute value of the difference between the absolute data uncertainty estimate from the PEN and that from the original model; therefore, we assume that $\mathbf{U_A}(\mathbf{x}';\Theta)$ in Eq. (11) can be ignored. Then expected total uncertainty in the original model can be estimated as follows.

$$\mathbf{U_T}(\mathbf{x}';\mathcal{W}) \approx \mathbb{H}[\tilde{\mathbf{y}'}|\mathbf{x}';\mathbf{\Theta}] - \mathbf{U_E}(\mathbf{x}';\Theta). \quad (12)$$

Assuming that $\mathbf{U_E}(\mathbf{x}';\Theta)$ reflects the magnitude of the entropy range in the original model for $\mathbf{x}'$, the entropy range of $\mathbf{f^w}(\mathbf{x}') \subset [0,1]$ can be approximated by incorporating the epistemic uncertainty of PEN from Eq. (12) as follows.

$$\Big[\mathbb{H}[\tilde{\mathbf{y}'}|\mathbf{x}';\mathbf{\Theta}] - \mathbf{U_E}(\mathbf{x}';\Theta) - \alpha\sqrt{\mathbf{U_E}(\mathbf{x}';\Theta)},$$
$$\mathbb{H}[\tilde{\mathbf{y}'}|\mathbf{x}';\mathbf{\Theta}] - \mathbf{U_E}(\mathbf{x}';\Theta) + \alpha\sqrt{\mathbf{U_E}(\mathbf{x}';\Theta)}\Big]. \quad (13)$$

This range can be rewritten as follows, by introducing empirical boundary constants b_L and b_R.

$$\Big[\mathbb{H}\big[\tilde{\mathbf{y}'}|\mathbf{x}';\Theta\big] - b_L\sqrt{\mathbf{U_E}(\mathbf{x}';\mathbf{\Theta})} - err,$$
$$\mathbb{H}\big[\tilde{\mathbf{y}'}|\mathbf{x}';\Theta\big] - b_R\sqrt{\mathbf{U_E}(\mathbf{x}';\mathbf{\Theta})} + err\Big], \quad (14)$$

where $b_L > b_R$, and err denotes the error bound. The parameters and error correction constraints associated with the estimated bounds of the entropy range were determined empirically. (Section 4.1)

3.5 Modeling Human Perception with Bayesian Agents

The neurobiological mechanisms of uncertainty processing are not fully known. However, evidence suggests that there is high stochasticity in the expression of uncertainty in the human brain (Ma et al. 2006; Fiser et al. 2010; Berkes et al. 2011; Orbán et al. 2016). Additionally, humans are good at reporting the level of their confidence. For example, they would report a low confidence value for the inputs that do not belong to any class.

We argue that the visual classification by human subjects can be modeled as a BNN through our experimental results. Our human experiments dealt with chest X-ray(CXR) images with binary labels (i.e, normal vs. abnormal). In our experiment, human subjects were presented with an image $\mathbf{x}$; then, they were asked to submit answers and the corresponding level of confidence at four different instances. Herein, we call this process *behavioral sampling*. A confidence level c places constraints on their choice probability of a subject as follows. For the binary classification problem, $\hat{\mathbf{y}} = \langle \frac{c}{2} + \frac{1}{2}, \frac{1}{2} - \frac{c}{2} \rangle$ when the first class is selected, and $\langle \frac{1}{2} - \frac{c}{2}, \frac{c}{2} + \frac{1}{2} \rangle$ when the second class is selected (where, $0 \leq c \leq 1$). Therefore, the value of the answer submitted by a subject for each image can be expressed as a two-dimensional probability vector from a softmax function. We also explored the possibility of modeling behavioral patterns with BNNs (Frenkel, Schrenk, and Martiniani 2017). As shown in Figure 1, we found that different responses were often reported for each sampling for the same $\mathbf{x}$ and subject. This enabled us to calculate the approximated uncertainty for image $\mathbf{x}$ of each subject. Interestingly, the higher the correspondence of the responses, the lower the total and epistemic uncertainties. Therefore, by introducing the human behavioral parameter $\mathbf{w}$ for the behavioral sampling of each subject, analogous to $\mathbf{w}$ in BNNs, we can consider a human subject as a Bayesian agent (original model), $\hat{\mathbf{y}} = \mathbf{f^w}(\mathbf{x})$, and approximate its behavior distribution.

3.6 Overview of the Proposed Framework

The proposed framework consists of a subject module (original model) and a PEN system (Figure 2). The subject module is a BNN model or a real human subject. In our behavioral experiments, a subject module indicated real human subjects. Prior to this main experiment, we also ran simulations in which the BNN-like model (human model agent, HMA) was used as a subject module (Section 4.3). The overall process can be defined as follows. In the subject module, sampling was performed for the image set $\mathbf{X}$ only once to obtain the training dataset $\widehat{\mathbf{D}}$ of the PEN. For the image set $\mathbf{X}'$, multiple samplings were independently performed 1000 and 4 times in the simulation and human behavior experiment, respectively, to obtain the test dataset $\widehat{\mathbf{D}'}$ of the PEN. The range of entropy is equal to the range of uncertainty; thus, we evaluated the performance of the PEN using a metric comparing its predicted range for $\mathbf{X}'$ and $\widehat{\mathbf{D}'}$.

4 Experimental Results

We verified the applicability of the PEN concept to simulations using several well-known datasets (Netzer et al. 2011; Xiao, Rasul, and Vollgraf 2017; Tschandl, Rosendahl, and Kittler 2018) to demonstrate that the PEN algorithm can infer the range and classification of the entropy reflecting the human visual decision uncertainty in real-world problems. As the applicability of the PEN was confirmed in our simulation experiment, a behavioral sampling experiment was conducted on human subjects. All collected data were included in the analyses. As a result of our experiment, PEN

	b_L	b_R	err
Human experiments	4	1	0.12-$\mathbf{U_E}(\mathbf{x}';\mathbf{\Theta}^*)$
Simulation experiments	4	1	0.20-$\mathbf{U_E}(\mathbf{x}';\mathbf{\Theta}^*)$

Table 1: Boundary constant and error.

Datasets	Image size	K	Number of PEN-training data	Number of PEN-test data	PUIA(%)	PUFIR(%)	CA(%)
SVHN	32×32×3	10	10000	1000	95.7±0.3	3.4±0.2	84.7±0.2
Fashion-MNIST	28×28×1	10	10000	1000	82.2±3.2	24.1±1.0	82.7±0.7
HAM10000	224×224×3	7	1000	515	95.2±0.6	22.2±1.1	81.6±1.7
CXR-A	224×224×1	2	220	50	80.2±4.6	19.4±2.9	92.8±1.4
CXR-B	224×224×1	2	220	50	75.8±5.3	20.0±2.3	79.8±7.8

Table 2: Results of simulation experiments with baseline.

	Scenario 1 / Scenario 2			Add 1 layer / Subtract 1 layer		
Datasets	PUIA(%)	PUFIR(%)	CA(%)	PUIA(%)	PUFIR(%)	CA(%)
SVHN	94.0 / 53.8	3.0 / 22.0	84.7 / 83.5	95.5 / 91.4	4.1 / 5.0	83.2 / 83.6
Fashion-MNIST	76.2 / 66.2	24.3 / 30.7	82.7 / 79.6	80.0 / 75.8	24.5 / 25.9	81.3 / 80.6
HAM10000	92.2 / 74.7	21.7 / 39.9	81.6 / 74.8	95.8 / 90.9	19.5 / 25.9	82.1 / 78.0
CXR-A	70.6 / 44.5	21.3 / 43.8	92.8 / 57.5	74.7 / 70.2	18.3 / 26.6	82.7 / 83.0
CXR-B	60.7 / 64.9	19.4 / 23.0	79.8 / 78.8	71.5 / 57.4	16.4 / 29.7	81.1 / 77.9

Table 3: Results of the ablation study. Scenario 1: Prediction performance in case of uncertainty estimation range for each data is applied equally in the estimation stage of the simulation. Scenario 2 : Prediction performance in case of random parameter sampling (MC dropout) is not applied at the output stage of HMA. Add 1 layer : Prediction performance when one layer was added from the baseline. Subtract 1 layer : Prediction performance when one layer was subtract from the baseline.

learned the predictive uncertainty distribution of the subject and predicted the uncertainty range and classification of the data to be sampled to the subject with high accuracy.

4.1 PEN Architecture and Training

Architecture. The generation of dataset $\widehat{\mathbf{D}}$ is contingent on the behavioral parameter. Therefore, the architecture of the PEN with the most appropriate capacity to fit $\widehat{\mathbf{D}}$ might be the same as the hypothetical model representing the human behavioral parameter probability space $\mathcal{W}$. Among the several ANN architectures that model the computational process of the human visual cortex, CORnet-Z (Kubilius et al. 2018) which is a simple feedforward CNN architecture, was adopted as the base architecture in this study. However, this does not mean that CORnet-Z is best suited for modeling the human visual decision uncertainty; further examination is required to explore more biologically plausible architectures.

Training. We used a 5 fold cross-validation ensemble approach, wherein we cross-divided the training dataset into training and validation datasets, and independently trained a network.

Inference boundary. The inference boundary constants used in Eq. (14) are summarized in Table 1.

4.2 Performance Metrics

The range of uncertainty for each data predicted by the PEN and the uncertainty of the subject calculated by the entropy are expressed between 0 and 1. The following metrics were defined to evaluate the performance of each subject.

Predictive Uncertainty Inference Accuracy (PUIA). It is defined as the percentage of reported uncertainty of the subject (HMA), which included in the PEN prediction range.

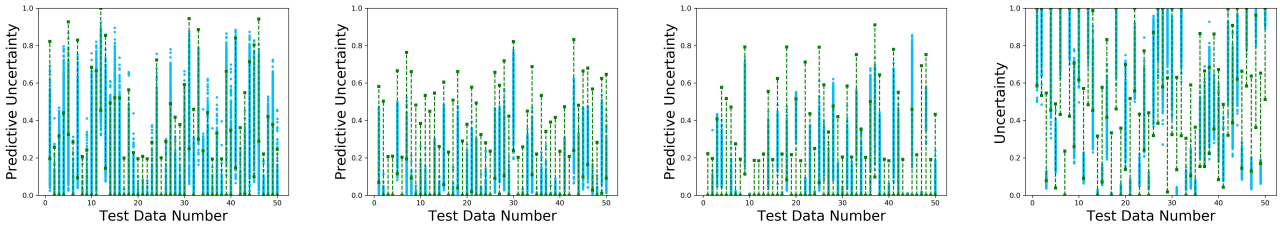
Predictive Uncertainty False Inference Rate (PUFIR). It is defined as the value of the percentage of predicted range, which does not overlap with the reported range of the subject (HMA) over total the entropy range [0,1].

Classification Accuracy (CA). It is defined as the percentage of correct prediction of PEN for the classification results (reported label) reported by the subject (HMA).

4.3 Simulation Study

The simulation involved sampling the behavior of a trained CNN model (HMA) instead of a human subject. The HMA generates output by applying an independent MC dropout to all independent input data in the training and test phase. For convenience, we divide each dataset into sub-datasets H, P, and T (Figure 2). The HMA was trained with H. From the trained HMA, PEN–training datasets were obtained by performing behavioral sampling once with P. PEN–test datasets were obtained by performing 1000 behavioral sampling with each image in T. We verified that the trained PEN predicted the uncertainty range and classification label by the HMA. CXR data consisted of H with separate images, other than those used in human experiments. As presented in Table 2, PEN demonstrated a high predictive performance in simulation experiments. In Figure 3, the predicted range of PEN and the actual output of HMA are shown for each dataset.

Ablation Study. To verify whether the epistemic uncertainty of the PEN is related to the magnitude of uncertainty range of the original model, the performance was evaluated using the average epistemic uncertainty for the total test dataset instead of each uncertainty value of the test datasets. This implies that Eq. (14) was calculated using $\frac{1}{N}\sum_{\mathbf{x}'\in\mathbf{X}'}\mathbf{U}_{\mathbf{E}}(\mathbf{x}';\mathbf{\Theta})$ instead of $\mathbf{U}_{\mathbf{E}}(\mathbf{x}';\mathbf{\Theta})$, where $\mathbf{X}'$ denotes the sets of N PEN-test datasets (Scenario 1). In addition, to verify the impact of MC dropout sampling when



(a) SVHN (b) Fashion MNIST (c) HAM10000 (d) CXR-A

Figure 3: Simulation results for each dataset (only 50 data is displayed). (Green dotted line: predictive uncertainty inference range, Blue dot: HMA output(1000 sampling) per each test data)

	Baseline			Add 1 layer		Subtract 1 layer	
	PUIA(%)	PUFIR(%)	CA(%)	PUIA(%)	CA(%)	PUIA(%)	CA(%)
CXR-A	74.7±12.9	50.4±11.3	70.6±7.9	80.8±7.9	67.5±8.8	32.8±28.2	68.2±8.7
CXR-B	79.5±20.1	46.6±13.2	64.6±9.3	82.4±6.6	62.7±9.1	51.0±39.4	65.0±9.1

Table 4: Summary of behavioral experiment results (For all 64 subjects). Shows the mean and standard deviation of the prediction performance of each trained PEN for all subjects. We compared performance of the PEN trained with a baseline architecture with those trained with an architecture in which one layer was added or subtracted from the baseline, respectively.

obtaining the PEN–training dataset we conducted an ablation study that applied the MC dropout to obtain the PEN–test dataset without applying the MC dropout to obtain the PEN–training dataset (Scenario 2). Both results exhibited poor PEN performance in comparison to the baseline versions (Table 3). These results demonstrate the relationship between the epistemic uncertainty of $\mathbf{x}'$ for PEN, the predictive uncertainty range of the original model, and the importance of MC sampling in this process.

Architecture Variation. To show that the most appropriate architecture of the PEN is the same as the HMA to be simulated, the performance of the PEN was compared by removing or adding one layer from the baseline. Here, baseline means that the architecture of the HMA and PEN are the same. The performance of the PEN with a smaller capacity than the baseline decreased significantly in all areas. In the simulation with an additional single layer, PUIA exhibited similar or slightly improved performance in comparison to the baseline, but CA demonstrated similar or slightly worse performance (Table 3).

4.4 Human Experiments

A CXR image was used in the behavior sampling experiment of human subjects. The CXR image contains a significant amount of information and is widely used in clinical situations, however, it is difficult to interpret owing to its high uncertainty (Pham et al. 2020). Two types of datasets were used in the experiment. CXR-A is labeled as normal or abnormal and CXR-B is labeled as edema or pneumonia diagnosis, and the two datasets are independent. The experimental dataset consisted of 270 images each. Each experimental dataset A and B consisted of 270 images that were divided into 220 and 50 images for the PEN–training and PEN–test datasets, respectively. The data for PEN training and test were randomly classified in advance, and the same settings were applied to all subjects. The experimental dataset was constructed by extracting images from the MIMIC-CXR dataset (Johnson et al. 2019a,b) and the CheXpert dataset (Irvin et al. 2019). To create the CXR datasets, labeled images were randomly chosen from the MIMIC-CXR and CheXpert datasets. A physician in our team then checked each image to discard all the incorrectly labeled images. As a result, the CXR-A and CXR-B datasets consist of 270 images in total.

All subjects were clinical physicians with an M.D and official government license. Human subject behavior sampling involved showing each image on a computer display with a size of 1024 $\times$ 1024 for 5 s. It required maximum concentration and received binary classification and confidence ranging 0 and 100. Images for the PEN test were sampled 4 times; the sampling for the same image was performed on different days to preclude memory interference. The subject ware prohibited from learning any chest X-ray images during the experiment period. To obtain the maximum accurate subjective confidence for each sampling, the subjects were informed beforehand that they will be awarded with high prizes in proportion to reporting high confidence with correct answers or low confidence with incorrect answers. In Figure 1, when the subjects reported that the uncertainty of the images was high in the PEN–test datasets, the classification responses for the images appeared to be inconsistent. Note that PEN-test datasets have been sampled multiple times, and the amount of each uncertainty can be calculated. It also appears that the epistemic uncertainty is associated with response mismatch, whereas the aleatoric uncertainty is less likely to be associated with response mismatch. This suggests the possibility that aleatoric uncertainty does not account for inconsistent choices.

Data Collection. We collected two sets of human behavior data, each consisting of the 64 physicians' medical diagnosis and their corresponding confidence values for each image of CXR-A and CXR-B dataset, respectively. Fifty images were sampled 4 times, making 840 responses per subject. The datasets also include a true label of each image, determined based on three votes from "Oracle" (radiologist). Although the current study does not make use of the true labels, we believe that this information would be useful for potential follow-up studies.

Summary of our results. The subjects had different prior knowledge related chest imaging; thus there was a difference between the estimated decision boundary and average uncertainty for chest imaging. The experimental results demonstrated that the PEN exhibits a highly accurate uncertainty prediction performance of an average of 75% or more for both experimental datasets (Table 4). The average CA for CXR-B was found to be rather low at 65% because the frequency of the inconsistent responses of the subjects that were repeatedly sampled for the PEN test is high. Unlike the simulation, the PUFIR exhibited a weak indicator because in human experiments, we sampled only 4 times for testing the PEN. The change in performance under the condition where a PEN layer was added or removed was similar to that in the simulation experiment. This provides implications for a fundamental approach to design neural networks architecture for human uncertainty modeling studies.

5 Conclusions

We proposed a new conceptual framework to understand human perceptual uncertainty from the perspective of BNNs and demonstrated that human uncertainty can be approximated through deterministic ensemble neural networks. The proposed model, called the PEN, can learn the expected value of human perceptual uncertainty as a function of behavioral parameters and predict the range of human uncertainty by efficiently inferring the uncertainty of the ensemble. This concept verified the practical applicability through large-scale medical image interpretation experiments by medical physicians, which involved real-world data with a wide range of perceptual uncertainty. Human uncertainty information is required for establishing an optimal learning strategy; however, a major limitation here is that this information can only be obtained by examining the samplings. Correspondingly, this study offered a new paradigm for uncertainty inference that efficiently resolves the difficulty of accessing human uncertainty information. In addition, our research is based on the premise that processing human uncertainty information and performing visual decision-making are similar to the process of the BNN.

Acknowledgments

We appreciate the 64 physicians participated in the human behavior experiments for this work. We thank Kumdori, the mascot of the Taejon International Exposition, Korea 1993, for giving us an inspiration for this work.

This work was supported by Smart Healthcare Based Thesis Research Grant through the Daewoong Foundation (DS183), Institute for Information & communications Technology Planning & Evaluation(IITP) grant funded by the Korea government(MSIT) (No.2019-0-01371, Development of brain-inspired AI with human-like intelligence; No. 2017-0-00451) and Samsung Research Funding Center of Samsung Electronics (No.SRFC-TC1603-06).

Ethics Statement

We used human data to fit our models. All the data were anonymized. The Institute Review Board of Korea Advanced Institute of Science and Technology (IRB number: KH2018-123) approved the study and experiments. All participants were informed of the potential risks due to experiments and submitted informed consent before the experiment. This work has the following potential impacts on society. 1) Expert recommendation system: Our system can assist experts in a variety of medical and commercial applications. However, the misuse of the system may lead to potential bias in people's judgments. 2) Education: It can be used to guide developing optimal learning strategies by distinguishing what the learners know and what they do not know. However, the failure of the system can confuse people. 3) AI engineering: It may be used to help people understand what machine learning algorithms learn from data. The misuse of the system may cause bias and security issues. 4) Human experiments: This system should not be used to influence ethical standards intentionally. The use of this system should be strictly limited to experts who not only follow a code of ethics, but also make objective and independent judgments (e.g., clinical physicians).

References

Barthelmé, S.; and Mamassian, P. 2009. Evaluation of objective uncertainty in the visual system. *PLoS computational biology* 5(9).

Berkes, P.; Orbán, G.; Lengyel, M.; and Fiser, J. 2011. Spontaneous cortical activity reveals hallmarks of an optimal internal model of the environment. *Science* 331(6013): 83–87.

Chai, L. R. 2018. *Uncertainty Estimation in Bayesian Neural Networks And Links to Interpretability*. Master's thesis, University of Cambridge.

Cohn, D. A.; Ghahramani, Z.; and Jordan, M. I. 1996. Active learning with statistical models. *Journal of artificial intelligence research* 4: 129–145.

Der Kiureghian, A.; and Ditlevsen, O. 2009. Aleatory or epistemic? Does it matter? *Structural safety* 31(2): 105–112.

Fiser, J.; Berkes, P.; Orbán, G.; and Lengyel, M. 2010. Statistically optimal perception and learning: from behavior to neural representations. *Trends in cognitive sciences* 14(3): 119–130.

Frenkel, D.; Schrenk, K. J.; and Martiniani, S. 2017. Monte Carlo sampling for stochastic weight functions. *Proceedings of the National Academy of Sciences* 114(27): 6924–6929.

Gal, Y. 2016. *Uncertainty in deep learning*. Ph.D. thesis, University of Cambridge.

Gal, Y.; and Ghahramani, Z. 2016. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, 1050–1059.

Gal, Y.; Islam, R.; and Ghahramani, Z. 2017. Deep bayesian active learning with image data. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, 1183–1192. JMLR. org.

Grinband, J.; Hirsch, J.; and Ferrera, V. P. 2006. A neural representation of categorization uncertainty in the human brain. *Neuron* 49(5): 757–763.

Hramov, A. E.; Frolov, N. S.; Maksimenko, V. A.; Makarov, V. V.; Koronovskii, A. A.; Garcia-Prieto, J.; Antón-Toro, L. F.; Maestú, F.; and Pisarchik, A. N. 2018. Artificial neural network detects human uncertainty. *Chaos: An Interdisciplinary Journal of Nonlinear Science* 28(3): 033607.

Hüllermeier, E.; and Waegeman, W. 2019. Aleatoric and epistemic uncertainty in machine learning: A tutorial introduction. *arXiv preprint arXiv:1910.09457* .

Irvin, J.; Rajpurkar, P.; Ko, M.; Yu, Y.; Ciurea-Ilcus, S.; Chute, C.; Marklund, H.; Haghgoo, B.; Ball, R.; Shpanskaya, K.; et al. 2019. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, 590–597.

Johnson, A.; Pollard, T.; Mark, R.; Berkowitz, S.; and Horng, S. 2019a. MIMIC-CXR Database (version 2.0.0). *PhysioNet. https://doi.org/10.13026/C2JT1Q.* .

Johnson, A. E.; Pollard, T. J.; Berkowitz, S. J.; Greenbaum, N. R.; Lungren, M. P.; Deng, C.-y.; Mark, R. G.; and Horng, S. 2019b. MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific Data* 6.

Kendall, A.; and Gal, Y. 2017. What uncertainties do we need in bayesian deep learning for computer vision? In *Advances in neural information processing systems*, 5574–5584.

Kingma, D. P.; Salimans, T.; and Welling, M. 2015. Variational dropout and the local reparameterization trick. In *Advances in neural information processing systems*, 2575–2583.

Kubilius, J.; Schrimpf, M.; Nayebi, A.; Bear, D.; Yamins, D. L.; and DiCarlo, J. J. 2018. Cornet: Modeling the neural mechanisms of core object recognition. *BioRxiv* 408385.

Lakshminarayanan, B.; Pritzel, A.; and Blundell, C. 2017. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in neural information processing systems*, 6402–6413.

Lee, S. W.; O'Doherty, J. P.; and Shimojo, S. 2015. Neural computations mediating one-shot learning in the human brain. *PLoS biology* 13(4).

Liu, J.; Paisley, J.; Kioumourtzoglou, M.-A.; and Coull, B. 2019. Accurate Uncertainty Estimation and Decomposition in Ensemble Learning. In *Advances in Neural Information Processing Systems*, 8950–8961.

Luce, R. 1959. Individual Choice Behavior: A theoretical analysis, New York, NY: John Willey and Sons.

Ma, W. J.; Beck, J. M.; Latham, P. E.; and Pouget, A. 2006. Bayesian inference with probabilistic population codes. *Nature neuroscience* 9(11): 1432–1438.

Malinin, A.; and Gales, M. 2018. Predictive uncertainty estimation via prior networks. In *Advances in Neural Information Processing Systems*, 7047–7058.

McClure, P. 2018. *Adapting deep neural networks as models of human visual perception*. Ph.D. thesis, University of Cambridge.

Nakada, M.; Chen, H.; and Terzopoulos, D. 2018. Deep learning of biomimetic visual perception for virtual humans. In *Proceedings of the 15th ACM Symposium on Applied Perception*, 1–8.

Netzer, Y.; Wang, T.; Coates, A.; Bissacco, A.; Wu, B.; and Ng, A. Y. 2011. Reading digits in natural images with unsupervised feature learning .

Orbán, G.; Berkes, P.; Fiser, J.; and Lengyel, M. 2016. Neural variability and sampling-based probabilistic representations in the visual cortex. *Neuron* 92(2): 530–543.

Peterson, J. C.; Battleday, R. M.; Griffiths, T. L.; and Russakovsky, O. 2019. Human uncertainty makes classification more robust. In *Proceedings of the IEEE International Conference on Computer Vision*, 9617–9626.

Pham, H. H.; Le, T. T.; Ngo, D. T.; Tran, D. Q.; and Nguyen, H. Q. 2020. Interpreting Chest X-rays via CNNs that Exploit Hierarchical Disease Dependencies and Uncertainty Labels. *arXiv preprint arXiv:2005.12734* .

Tong, S. 2001. *Active Learning: Theory and Applications*. Ph.D. thesis, Stanford University.

Tschandl, P.; Rosendahl, C.; and Kittler, H. 2018. The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific data* 5: 180161.

van Bergen, R. S.; and Jehee, J. F. 2019. Probabilistic representation in human visual cortex reflects uncertainty in serial decisions. *Journal of Neuroscience* 39(41): 8164–8176.

Wang, G.; Li, W.; Aertsen, M.; Deprest, J.; Ourselin, S.; and Vercauteren, T. 2018. Test-time augmentation with uncertainty estimation for deep learning-based medical image segmentation. In *t Conference on Medical Imaging with Deep Learning*.

Xiao, H.; Rasul, K.; and Vollgraf, R. 2017. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747* .

Yamins, D. L.; Hong, H.; Cadieu, C.; and DiCarlo, J. J. 2013. Hierarchical modular optimization of convolutional networks achieves representations similar to macaque IT and human ventral stream. In *Advances in neural information processing systems*, 3093–3101.

Yamins, D. L.; Hong, H.; Cadieu, C. F.; Solomon, E. A.; Seibert, D.; and DiCarlo, J. J. 2014. Performance-optimized hierarchical models predict neural responses in higher visual

cortex. *Proceedings of the National Academy of Sciences* 111(23): 8619–8624.