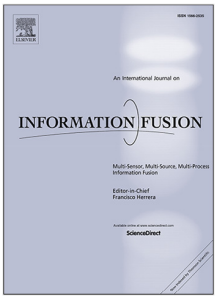
Journal Pre-proof



A review of uncertainty quantification in deep learning: Techniques, applications and challenges

Moloud Abdar, Farhad Pourpanah, Sadiq Hussain, Dana Rezazadegan, Li Liu, Mohammad Ghavamzadeh, Paul Fieguth, Xiaochun Cao, Abbas Khosravi, U. Rajendra Acharya, Vladimir Makarenkov, Saeid Nahavandi

PII: S1566-2535(21)00108-1
DOI: https://doi.org/10.1016/j.inffus.2021.05.008
Reference: INFFUS 1411

To appear in: *Information Fusion*

Received date : 13 January 2021
Revised date : 20 April 2021
Accepted date : 15 May 2021

Please cite this article as: M. Abdar, F. Pourpanah, S. Hussain et al., A review of uncertainty quantification in deep learning: Techniques, applications and challenges, *Information Fusion* (2021), doi: https://doi.org/10.1016/j.inffus.2021.05.008.

A Review of Uncertainty Quantification in Deep Learning: Techniques, Applications and Challenges

Moloud Abdar[a,*], Farhad Pourpanah[b], Sadiq Hussain[c], Dana Rezazadegan[d], Li Liu[e], Mohammad Ghavamzadeh[f], Paul Fieguth[g], Xiaochun Cao[h], Abbas Khosravi[a], U Rajendra Acharya[i,j,k], Vladimir Makarenkov[l], Saeid Nahavandi[a]

[a]*Institute for Intelligent Systems Research and Innovation (IISRI), Deakin University, Geelong, Australia*
[b]*College of Mathematics and Statistics, Guangdong Key Lab. of Intelligent Information Processing, Shenzhen University, China*
[c]*System Administrator, Dibrugarh University, Dibrugarh, India*
[d]*Department of Computer Science and Software Engineering, Swinburne University of Technology, Melbourne, Australia*
[e]*Center for Machine Vision and Signal Analysis, University of Oulu, Oulu, Finland*
[f]*Google Research, Google, USA*
[g]*Department of Systems Design Engineering, University of Waterloo, Waterloo, Canada*
[h]*State Key Laboratory of Information Security, Institute of Information Engineering, Chinese Academy of Sciences, Beijing, China*
[i]*Department of Electronics and Computer Engineering, Ngee Ann Polytechnic, Clementi, Singapore*
[j]*Department of Biomedical Engineering, School of Science and Technology, SUSS University, Singapore*
[k]*Department of Biomedical Informatics and Medical Engineering, Asia University, Taichung, Taiwan*
[l]*Department of Computer Science, University of Quebec in Montreal, Montreal (QC), Canada*

Abstract

Uncertainty quantification (UQ) methods play a pivotal role in reducing the impact of uncertainties during both optimization and decision making processes. They have been applied to solve a variety of real-world applications in science and engineering. Bayesian approximation and ensemble learning techniques are two of the most widely-used types of uncertainty quantification (UQ) methods. In this regard, researchers have proposed different UQ methods and examined their performance in a variety of applications such as computer vision (e.g., self-driving cars and object detection), image processing (e.g., image restoration), medical image analysis (e.g., medical image classification and segmentation), natural language processing (e.g., text classification, social media texts and recidivism risk-scoring), bioinformatics, etc. This study reviews recent advances in UQ methods used in deep learning, investigates the application of these methods in reinforcement learning, and highlights the fundamental research challenges and directions associated with the UQ field.

Keywords: Artificial intelligence, Uncertainty quantification, Deep learning, Machine learning, Bayesian statistics, Ensemble learning.

1. Introduction

Everyday scenarios deal with uncertainties in a variety fields, from investment opportunities and medical diagnoses to sports games and weather forecasting, and in all cases, with the objective to make decisions based on collected observations and uncertain-domain knowledge. Models developed using machine learning and deep learning are widely used for all types of inference and decision making, meaning that it is increasingly important to evaluate the reliability and efficacy of artificial intelligence (AI) systems before they could be applied in practice [1], since the predictions made by such models are subject to noise and model inference

*Corresponding author
Email addresses: m.abdar1987@gmail.com, mabdar@deakin.edu.au (Moloud Abdar), farhad.086@gmail.com (Farhad Pourpanah), sadiq@dibru.ac.in (Sadiq Hussain), drezazadegan@swin.edu.au (Dana Rezazadegan), li.liu@oulu.fi (Li Liu), ghavamza@google.com (Mohammad Ghavamzadeh), pfieguth@uwaterloo.ca (Paul Fieguth), caoxiaochun@iie.ac.cn (Xiaochun Cao), abbas.khosravi@deakin.edu.au (Abbas Khosravi), aru@np.edu.sg (U Rajendra Acharya), makarenkov.vladimir@uqam.ca (Vladimir Makarenkov), saeid.nahavandi@deakin.edu.au (Saeid Nahavandi)

errors [2]. It is thus highly desirable to represent uncertainty in a trustworthy manner in any AI-based system. The principles of uncertainty play an important role in AI settings such as concrete learning algorithms [3] and active learning (AL) [4, 5].

Sources of uncertainty arise when the test and training data are mismatched, and data uncertainty occurs because of class overlap or due to the presence of noise in the data [6]; however, estimating knowledge uncertainty is significantly more difficult than estimating data uncertainty [7]. There are two main sources of uncertainty, aleatoric and epistemic uncertainty [8] (see Fig. 1).

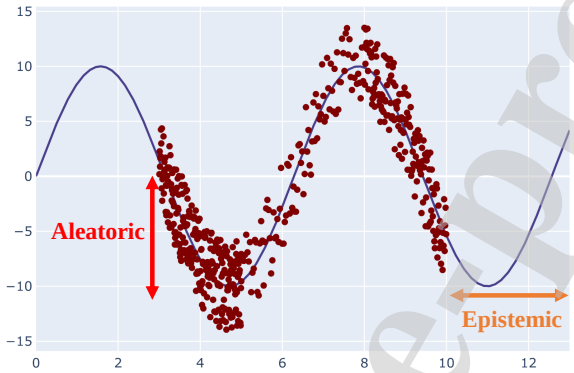


Fig. 1: A schematic view of the main differences between aleatoric and epistemic uncertainty.

The irreducible uncertainty in data that gives rise to uncertainty in predictions is aleatoric uncertainty (also known as data uncertainty). This type of uncertainty is not a property of the model, but rather is an inherent property of the data distribution, and hence, it is irreducible. In contrast, epistemic uncertainty (also known as knowledge uncertainty) occurs due to inadequate knowledge. One can define models to answer different questions in model-based prediction. For data-rich problems, there may be massive collections of data that are information poor [10]. In such cases, AI-based methods can be used to define the efficient models that characterize the emergent features of the data. Very often, these data are incomplete, noisy, discordant or multimodal [2].

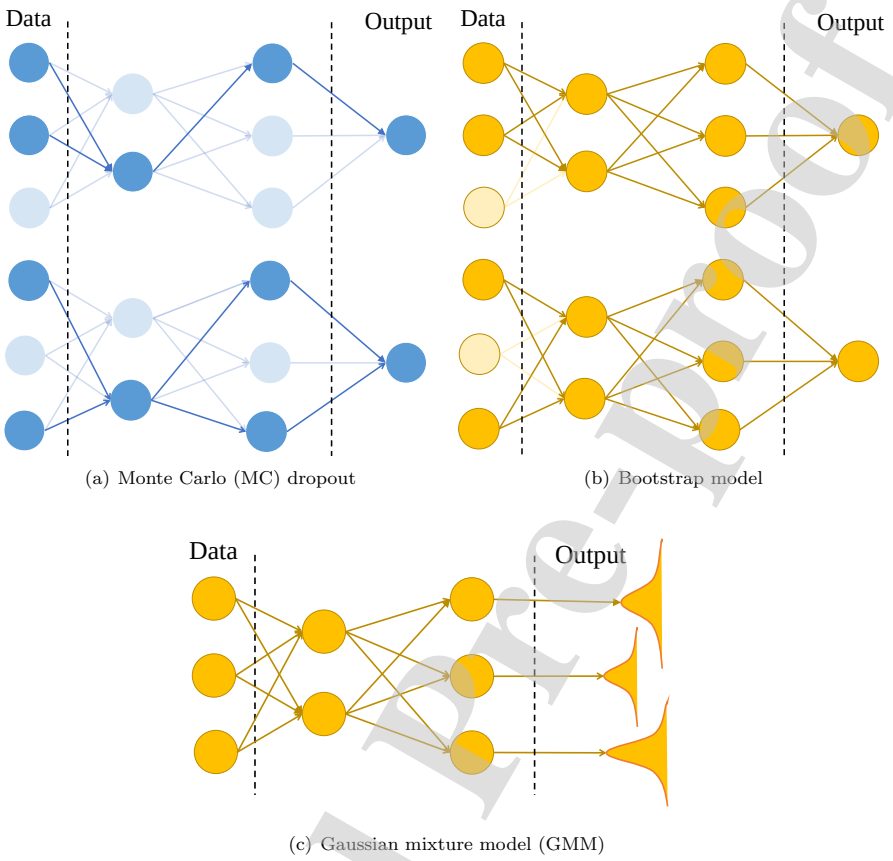
Uncertainty quantification (UQ) currently underpins many critical decisions, and predictions made without UQ are usually not trustworthy and accurate. To understand the deep learning (DL) [11, 12] process life cycle, we need to comprehend the role of UQ in DL. DL models start with a collection of the most comprehensive and potentially relevant datasets available for the decision making process. The DL scenario is then designed to meet some performance goals to select the most appropriate DL architecture, and then a model is trained based on the data. The iterative training process optimizes different learning parameters, which are adjusted until the network provides a satisfactory level of performance.

There are several uncertainties that need to be quantified in the steps involved:

(i) the selection and collection of the training data,
(ii) the completeness and accuracy of the training data,
(iii) an understanding of the DL (or traditional machine learning) model with performance bounds and limitations, and
(iv) uncertainties corresponding to the performance of the model based on operational data [13].

Next, data-driven approaches, particularly those similar to DL, pose at least four overlapping challenges:

(i) an absence of theory,
(ii) an absence of causal models,
(iii) sensitivity to imperfect data, and
(iv) computational costs.



(a) Monte Carlo (MC) dropout (b) Bootstrap model

(c) Gaussian mixture model (GMM)

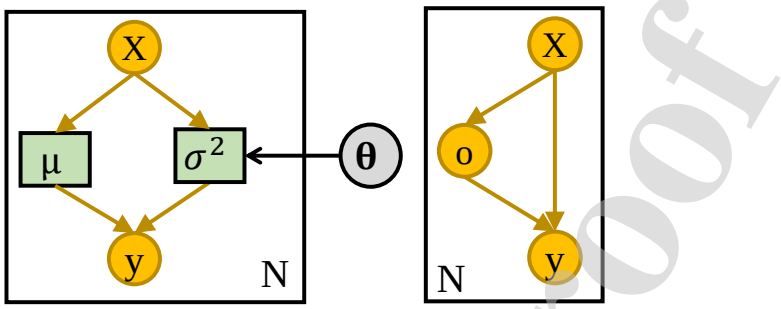
Fig. 2: Schematic view of three different uncertainty models with the related network architectures, reproduced based on [9].

Uncertainty estimation and quantification have been studied extensively in DL and traditional machine learning, and a schematic comparison of the three different uncertainty models [9] (MC dropout, the Bootstrap model and the GMM) is provided in Fig. 2. In addition, two graphical representations of uncertainty-aware models (BNN and OoD) are illustrated in Fig. 3. In what follows, we provide a brief summary of recent studies that examine the effectiveness of methods of dealing with uncertainties.

1.1. Research Objectives and Outline

In the era of big data, ML and DL, intelligent use of various raw data has great potential to benefit a wide variety of areas. However, UQ in different ML and DL methods can significantly increase the reliability of their results. Ning et al. [15] summarized and classified the main contributions of the data-driven optimization paradigm under uncertainty. This paper reviewed data-driven optimization only. In another study, Kabir et al. [16] reviewed neural network-based UQ. The authors focused on probabilistic forecasting and prediction intervals (PIs) as they are among most widely used techniques in the literature for UQ.

We note that, from 2010 to 2020 (end of June), more than 2500 papers on UQ in AI were published in various fields (e.g., computer vision, image processing, medical image analysis, signal processing, and natural language processing). On the one hand, we ignore a large number of papers due to a lack of adequate connection with the subject of our review. On the other hand, although many papers that we reviewed have been published in related conferences and journals, many papers were found on open-access repositories as electronic preprints (i.e., arXiv), and we reviewed them due to their high quality and full



(a) BNN (b) OoD classifier

Fig. 3: A graphical representation of two different uncertainty-aware (UA) models, reproduced based on [14].

relevance to the subject. We aimed to cover most of the related articles in this review paper. This review can therefore serve as a comprehensive guide to the reader in navigating this fast-growing research field.

Unlike previous review papers in the field of UQ, this study reviews most recent articles published on quantifying uncertainty in AI (ML and DL) using different approaches. In addition, we are keen to discover how UQ can impact real cases and address uncertainty in AI to help obtain reliable results. Moreover, finding important conversations on existing methods is an excellent way to reveal paths to future research. In this regard, this review paper can provide more input to future researchers working on UQ in ML and DL. We investigate more recent studies in the domain of UQ applied to ML and DL methods. Therefore, we summarize few existing studies on UQ in ML and DL. It is worth mentioning that the main purpose of this study is not to compare the performance of the different UQ methods proposed because these methods are introduced for different data and specific tasks. For this reason, we argue that comparing the performance of all methods is beyond the scope of this study. For this reason, this study mainly focuses on important areas including DL, ML and reinforcement learning (RL). Hence, the main contributions of this study are as follows:

- To the best of our knowledge, this is the first comprehensive review paper regarding UQ methods used in ML and DL methods, which is valuable for researchers in this domain.
- A comprehensive review of newly proposed UQ methods is provided.
- The main categories of important applications of UQ methods are also listed.
- The main research gaps of UQ methods are pointed out.
- Some solid future research directions are discussed.

2. Preliminaries

In this section, we explain the structure of feedforward neural networks (NNs) and Bayesian modeling to discuss uncertainty in detail.

2.1. Feedforward Neural Networks

In this section, the structure of a single-hidden-layer neural network (NN) [17] is explained, and this can be extended to multiple layers. Suppose x is a D-dimensional input vector; we use a linear map W_1 and bias b to transform x into a row vector with Q elements, i.e., $W_1x + b$. Next, a nonlinear transfer function $\sigma(.)$, such as the rectified linear unit (ReLU), can be applied to obtain the output of the hidden layer. Then, another linear function W_2 can be used to map the hidden layer to the output:

$$\hat{y} = \sigma(xW_1 + b)W_2. \tag{1}$$

For classification, to compute the probability of X belonging to a label c in the set $\{1, ..., C\}$, the normalized score is obtained by passing the model output $\hat{y}$ through a softmax function $\hat{p}_d = \exp(\hat{y}_d)/(\sum_{d'} \exp(\hat{y}_{d'}))$. Then, softmax loss is used:

$$E^{W_1,W_2,b}(X,Y) = -\frac{1}{N}\sum_{i=1}^{N} \log(\hat{p}_{i,c_i}), \tag{2}$$

where $X = (x_1, ..., x_N)$ and $Y = (y_1, ..., y_N)$ are the inputs and their corresponding outputs, respectively. For regression, the Euclidean loss can be used:

$$E^{W_1,W_2,b}(X,Y) = \frac{1}{2N}\sum_{i=1}^{N} \|y_i - \hat{y}\|^2. \tag{3}$$

2.2. Uncertainty Modeling

As mentioned above, there are two main types of uncertainty: epistemic (model uncertainty) and aleatoric (data uncertainty) [18]. Aleatoric uncertainty has two types: *homoscedastic* and *heteroscedastic* [19].

Predictive uncertainty (PU) consists of two parts, *(i)* epistemic uncertainty (EU) and (ii) aleatoric uncertainty (AU), and it can be written as the sum of these two parts:

$$PU = EU + AU. \tag{4}$$

Epistemic uncertainty can be formulated as a probability distribution over the model parameters. Let $D_{tr} = \{X, Y\} = \{(x_i, y_i)\}_{i=1}^{N}$ denote a training dataset with inputs $x_i \in \Re^D$ and their corresponding classes $y_i \in \{1, ..., C\}$, where C represents the number of classes. The aim is to optimize the parameters, i.e., ω, of a function $y = f^{\omega}(x)$ that can produce the desired output. To achieve this, the Bayesian approach defines a model likelihood, i.e., $p(y|x, \omega)$. For classification, softmax likelihood can be used:

$$p(y = c|x, \omega) = \frac{\exp(f_c^{\omega}(x))}{\sum_{c'} \exp(f_{c'}^{\omega}(x))}. \tag{5}$$

and Gaussian likelihood can be assumed for regression:

$$p(y|x, \omega) = \mathcal{N}(y; f^{\omega}(x), \tau^{-1}I), \tag{6}$$

where τ represents the model precision.

The posterior distribution, i.e., $p(\omega|x, y)$, for a given dataset D_{tr} over ω obtained by applying Bayes' theorem can be written as follows:

$$p(\omega|X, Y) = \frac{p(Y|X, \omega)p(\omega))}{p(Y|X)}. \tag{7}$$

For a given test sample x^*, a class label with regard to $p(\omega|X, Y)$ can be predicted:

$$p(y^*|x^*, X, Y) = \int p(y^*|x^*, \omega)p(\omega|X, Y)d\omega. \tag{8}$$

This process is called inference or marginalization. However, $p(\omega|X, Y)$ cannot be computed analytically but can be approximated by variational parameters, i.e., $q_\theta(\omega)$. The aim is to approximate a distribution that is close to the posterior distribution obtained by the model. As such, the Kullback-Leibler (KL) [20] divergence needs to be minimized with regard to θ. The level of similarity among two distributions can be measured as follows:

$$KL(q_\theta(\omega)\|p(\omega|X, Y)) = \int q_\theta(\omega) \log \frac{q_\theta(\omega)}{p(\omega|X, Y)} d\omega. \tag{9}$$

The predictive distribution can be approximated by minimizing the KL divergence, as follows:

$$p(y^*|x^*, X, Y) \approx \int p(y^*|x^*, \omega) q_\theta^*(\omega) d\omega =: q_\theta^*(y^*, x^*), \tag{10}$$

where $q_\theta^*(\omega)$ indicates the optimized objective.

KL divergence minimization can also be rearranged into the *evidence lower bound* (ELBO) maximization [21]:

$$\mathcal{L}_{VI}(\theta) := \int q_\theta(\omega) \log p(Y|X, \omega) d\omega - KL(q_\theta(\omega) \| p(\omega)), \tag{11}$$

where $q_\theta(\omega)$ can describe the data well by maximizing the first term and can approach the prior as closely as possible by minimizing the second term. This process is called variational inference (VI). Dropout VI is one of the most common approaches and has been widely used to approximate inference in complex models [22]. The minimization objective is as follows [23]:

$$\mathcal{L}(\theta, p) = -\frac{1}{N} \sum_{i=1}^{N} \log p(y_i|x_i, \omega) + \frac{1-p}{2N} \|\theta\|^2, \tag{12}$$

where N and p represent the number of samples and the dropout probability, respectively.

To obtain the data-dependent uncertainty, the precision τ in (6) can be formulated as a function of the data. One approach to obtaining the epistemic uncertainty is to mix two functions—the predictive mean, i.e., $f^\theta(x)$, and the model precision, i.e., $g^\theta(x)$—and the likelihood function can then be written as $y_i = \mathcal{N}(f^\theta(x), g^\theta(x)^{-1})$. A prior distribution is placed on the weights of the model, and then the amount of change in the weights for the given data samples is computed. The Euclidian distance loss function (3) can be adapted as follows:

$$\begin{aligned} E^{W_1,W_2,b} := & \\ & \frac{1}{2}(y - f^{W_1,W_2,b}(x)) g^{W_1,W_2,b}(x) (y - f^{W_1,W_2,b}(x))^T - \frac{1}{2} \log det\ g^{W_1,W_2,b} + \frac{D}{2} \log 2\pi \\ & = -\log \mathcal{N}(f^\theta(x), g^\theta(x)^{-1}). \end{aligned} \tag{13}$$

The predictive variance can be obtained as follows:

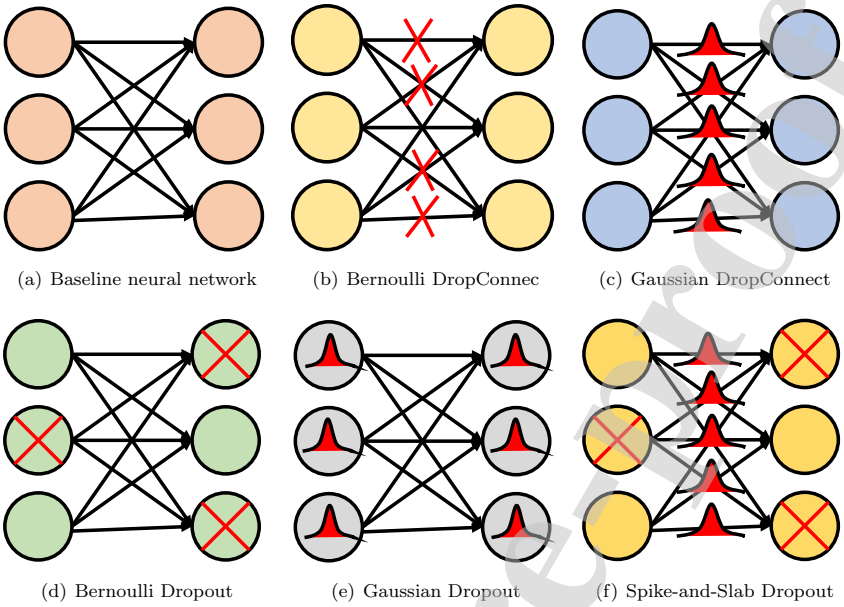
$$\widehat{Var}[x^*] := \frac{1}{T} \sum_{t=1}^{T} g^{\widetilde{\omega_t}}(x) \mathbf{I} + f^{\widetilde{\omega_t}}(x^*)^T f^{\widetilde{\omega_t}}(x^*) - \widetilde{\mathbb{E}}[y^*]^T \widetilde{\mathbb{E}}[y^*] \underset{T \to \infty}{\longrightarrow} Var_{q_\theta^*(y^*|x^*)}[y^*]. \tag{14}$$

3. Uncertainty Quantification Using Bayesian Techniques

Despite the success of standard DL methods in solving various real-word problems, they cannot provide information about the reliability of their predictions. To address this issue, Bayesian deep learning (BDL) and Bayesian NNs (BNNs) [24, 25, 26, 27, 28, 29, 30, 31] can be used to interpret the model parameters. BNNs and BDL are robust to overfitting problems and can be trained on both small and large (big) datasets [32].

3.1. Monte Carlo (MC) Dropout

As stated earlier, it is difficult to compute an exact posterior inference, but it can be approximated. In this regard, Monte Carlo (MC) [33] is an effective method. However, it is a slow and computationally expensive method when integrated into a deep architecture. To combat this, MC dropout was introduced; it uses dropout [34] as a regularization term to compute the prediction uncertainty [35]. Dropout is an effective technique that has been widely used to solve overfitting problems in deep NNs (DNNs). During the training process, dropout randomly drops some units of the NN to prevent excessive co-tuning. Assume an NN with L layers, where W_l, b_l and K_l denote the weight matrices, bias vectors and dimensions of the lth



(a) Baseline neural network (b) Bernoulli DropConnec (c) Gaussian DropConnect

(d) Bernoulli Dropout (e) Gaussian Dropout (f) Spike-and-Slab Dropout

Fig. 4: A graphical representation of several visualizations of variational distributions on a simple NN, which is reproduced based on [36].

layer, respectively. The output of the NN and the target class of the ith input x_i $(i = 1, ..., N)$ are indicated by $\hat{y}_i$ and y_i, respectively. The objective function using L_2 regularization can be written as:

$$\mathcal{L}_{dropout} := \frac{1}{N}\sum_{i=1}^{N} E(y_i, \hat{y_i}) + \lambda \sum_{l=1}^{L}(\|W_i\|_2^2 + \|b_i\|_2^2). \tag{15}$$

Dropout samples binary variables for each input data value and every network unit in each layer (except the output layer), with a probability of p_i for the ith layer; if its value is 0, the unit i is dropped for a given input data value. The same values are used in the backward pass to update the parameters. Fig. 4 shows several visualizations of variational distributions on a simple NN [36].

Several studies have used MC dropout [37] to estimate the UQ. Wang et al. [38] analyzed epistemic and aleatoric uncertainty for deep convolutional NN (CNN)-based medical image segmentation problems at both pixel and structure levels. They augmented the input image during the test phase to estimate the transformation uncertainty. Specifically, MC sampling was used to estimate the distribution of the output segmentation. Liu et al. [39] proposed a unified model using SGD to approximate both epistemic and aleatoric uncertainty for CNNs in the presence of universal adversarial perturbations. The epistemic uncertainty was estimated by applying MC dropout with Bernoulli distribution at the output of the neurons. In addition, they introduced texture bias to better approximate the aleatoric uncertainty. Nasir et al. [40] used MC dropout to estimate four types of uncertainties, including the variance of MC samples, predictive entropy, and mutual information (MI), in a 3D CNN to segment lesions from MRI sequences.

In [41], two dropout methods, i.e., elementwise Bernoulli dropout [34] and spatial Bernoulli dropout [42], were implemented to compute the model uncertainty in BNNs for end-to-end autonomous vehicle control. McClure and Kriegeskorte [36] expressed that the sampling of weights using Bernoulli or Gaussian dropout can lead to a more accurate depiction of uncertainty than the sampling of units. However, according to the outcomes obtained in [36], it can be argued that using either Bernoulli or Gaussian dropout can improve the classification accuracy of a CNN. Based on these findings, a novel model (called spike-and-slab sampling) was proposed by combining Bernoulli and Gaussian dropout.

Do et al. [61] modified U-Net [62], which is a CNN-based deep model, to segment myocardial arterial

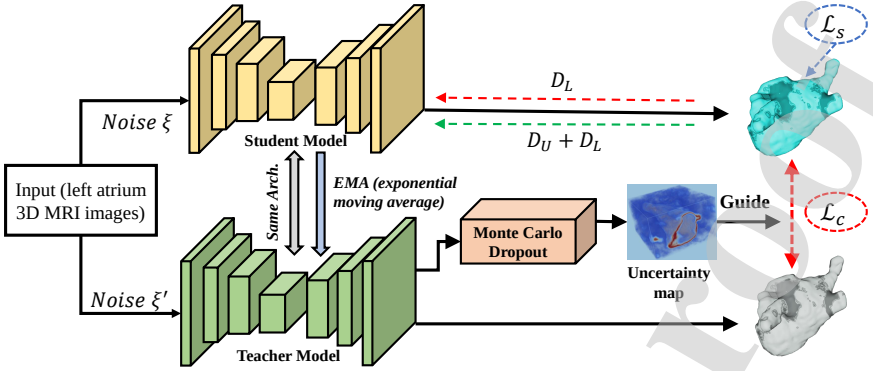


Fig. 5: A general view demonstrating the application of the semi-supervised UA-MT framework to LA segmentation, which is reproduced based on [43].

Table 1: A summary of studies that applied the original MC dropout to approximate uncertainty along with their applications (sorted by year).

Study	Year	Method	Application	Code
Kendal et al. [44]	2015	SegNet [45]	semantic segmentation	√
Leibig et al. [46]	2017	CNN	diabetic retinopathy	√
Choi et al. [47]	2017	mixture density network (MDN) [48]	regression	×
Jung et al. [49]	2018	full-resolution ResNet [50]	brain tumor segmentation	×
Wickstrom et al. [51]	2018	FCN [52] and SehNet [45]	polyps segmentation	×
Jungo et al. [53]	2018	FCN	brain tumor segmentation	×
Vandal et al. [54]	2018	Variational LSTM	predict flight delays	×
Devries and Taylor [55]	2018	CNN	medical image segmentation	×
Tousignant et al. [56]	2019	CNN	MRI images	×
Norouzi et al. [57]	2019	FCN	MRI images segmentation	×
Roy et al. [58]	2019	Bayesian FCNN	brain images (MRI) segmentation	√
Filos et al. [59]	2019	CNN	diabetic retinopathy	√
Harper and Southern [60]	2020	RNN and CNN	emotion prediction	×

spin labeling and estimate uncertainty. Specifically, batch normalization and dropout are added after each convolutional layer and resolution scale, respectively. Later, Teye et al. [63] proposed MC batch normalization (MCBN), which can be used to estimate the uncertainty of networks with batch normalization. They showed that batch normalization can be considered an approximate Bayesian model. Yu et al. [43] proposed a semi-supervised model to segment the left atrium from 3D MR images. This model consists of two modules, the teacher and student, and they are used in a UA framework called the UA self-ensembling mean teacher (UA-MT) model (see Fig. 5). As such, the student model learns from the teacher model by minimizing the segmentation and consistency losses of the labeled samples and targets of the teacher model, respectively. In addition, the UA framework based on MC dropout was designed to help the student model learn a better model by using uncertainty information obtained from the teacher model. Table 1 lists studies that directly applied MC dropout to approximate uncertainty along with their applications.

3.1.1. Comparison of MC Dropout with Other UQ Methods

Recently, several studies have been conducted to compare different UQ methods. For example, Foong et al. [65] empirically and theoretically studied MC dropout and mean-field Gaussian VI. They found that both models can express uncertainty well in shallow BNNs. However, mean-field Gaussian VI could not

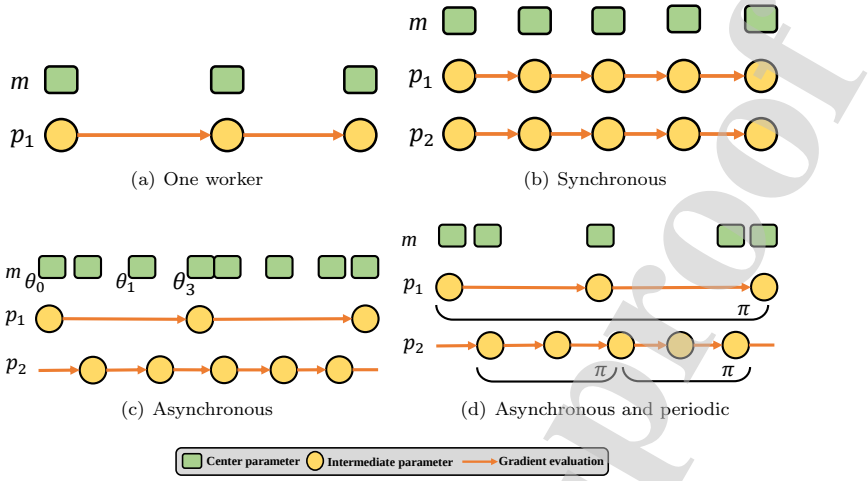


Fig. 6: Graphical implementations of different SG-MCMC models, which are reproduced based on [64].

approximate the posterior well in estimating uncertainty for deep BNNs. Ng et al. [66] compared MC dropout with BBB using U-Net [62] as a base classifier. Siddhant et al. [67] empirically studied various DAL models for natural language processing (NLP). During prediction, they applied dropout to CNNs and recurrent NNs (RNNs) to estimate the uncertainty. Hubschneider et al. [9] compared MC dropout with a bootstrap ensembling-based method and a Gaussian mixture for the task of vehicle control. In addition, Mukhoti [68] applied MC dropout in several models to estimate uncertainty in regression problems. Kennamer et al. [69] empirically studied MC dropout under astronomical observing conditions.

3.2. Markov Chain Monte Carlo (MCMC)

Markov chain Monte Carlo (MCMC) [70] is another effective method that has been used to approximate inference. It starts by taking a randomly drawn value z_0 from the distribution $q(z_0)$ or $q(z_0|x)$. Then, it applies a stochastic transition to z_0, as follows:

$$Z_t \sim q(z_t|z_{t-1}, x). \tag{16}$$

This transition operator is chosen and repeated T times, and the outcome, which is a random variable, converges in the distribution to the exact posterior. Salakhutdinov et al. [71] used MCMC to approximate a predictive distribution of the rating values of movies. Despite the success of conventional MCMC, the sufficient number of iterations is unknown. In addition, MCMC requires much time to converge to a desired distribution [33]. Several studies have been conducted to overcome these shortcomings. For example, Salimans et al. [72] expanded the space to a set of *auxiliary random variables* and interpreted the stochastic Markov chain as a variational approximation.

Stochastic gradient MCMC (SG-MCMC) [73, 74] was proposed to train DNNs. It needs to estimate only the gradient on small sets of mini-batches. In addition, SG-MCMC can converge to the true posterior by decreasing the step size [75, 76]. Gong et al. [77] combined amortized inference with SG-MCMC to increase the generalization ability of the model. Li et al. [64] proposed an accelerating SG-MCMC to improve the speed of the conventional SG-MCMC (see Fig. 6 for implementations of different SG-MCMC models). However, within a short time, SG-MCMC suffers from a bounded estimation error [78], and it loses the surface when applied to multilayer networks [79]. In this regard, Zhang et al. [80] developed a cyclical SG-MCMC (cSG-MCMC) to compute the posterior over the weights of NNs. Specifically, a cyclical step size was used instead of a decreasing one. A large step size allows the sampler to make large moves, while a small step size encourages the sampler to explore local modes.

Although SG-MCMC reduces the computational complexity by using a small subset, i.e., a mini-batch, of the dataset at each iteration to update the model parameters, these small subsets of data add noise to

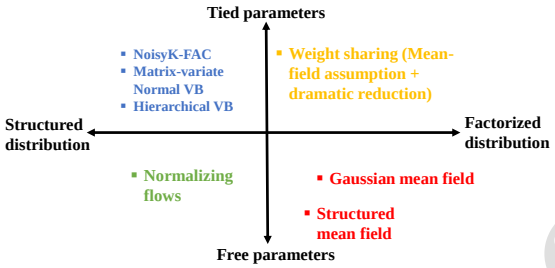


Fig. 7: A summary of various VI methods for BDL, which is reproduced based on [89]. Note that *Weight sharing* (*Mean − field assumption + dramatic reduction*) is added based on the method proposed in [89].

the model and consequently increase the uncertainty of the system. To address this issue, Luo et al. [81] introduced a sampling method called *thermostat-assisted continuously tempered Hamiltonian Monte Carlo*, which is an extended version of conventional Hamiltonian MC (HMC) [82]. Note that HMC is an MCMC method [83]. Specifically, they used Nosé-Hoover thermostats [84, 85] to handle the noise generated by the mini-batch datasets. Later, dropout HMC (D-HMC) [83] was proposed for uncertainty estimation, and it was compared with SG-MCMC [73] and SGLD [86].

In addition, MCMC was integrated into generative-based methods to approximate posteriors. For example, in [87], MCMC was applied to stochastic object models, which were learned by generative adversarial networks (GANs), to approximate the ideal observer. In [88], a visual tracking system based on a variational autoencoder (VAE) MCMC (VAE-MCMC) was proposed.

3.3. Variational Inference (VI)

Variational inference (VI) is an approximation method that learns the posterior distribution over BNN weights. VI-based methods consider the Bayesian inference problem as an optimization problem that is used by the SGD to train DNNs. Fig. 7 summarizes various VI methods for BNNs [89].

For BNNs, VI-based methods aim to approximate posterior distributions over the weights of the NN. To achieve this, the loss can be defined as follows:

$$\mathcal{L}(\Phi) \approx \frac{1}{2|\mathcal{D}|}\sum_{i=1}^{|\mathcal{D}|}\mathcal{L}_R(y^{(i)}, x^{(i)}) + \frac{1}{|\mathcal{D}|}KL(q_\phi(w)\|p(w)), \tag{17}$$

where $|\mathcal{D}|$ indicates the number of samples, and

$$\mathcal{L}_R(y, x) = -\log(\hat{\tau}_x)^T 1 + \|\sqrt{\hat{\tau}}_x \odot (y - \hat{\mu}_x)\|^2 \tag{18}$$

$$\hat{\mu}_x = \hat{\mu}(x, w_\mu); \qquad w \sim q_\phi(w) \tag{19}$$

$$\hat{\tau}_x = \hat{\tau}(x, w_r). \tag{20}$$

where $\odot$ and 1 represent the element-wise product and a vector filled with ones, respectively. Eq. (17) can be used to compute (10).

Posch et al. [90] defined the variational distribution using a product of Gaussian distributions along with diagonal covariance matrices. For each network layer, a posterior uncertainty of the network parameter was represented. Later, in [91], the diagonal covariance matrices were replaced with the traditional ones to allow the network parameters to correlate with each other. Inspired by transfer learning and empirical Bayes (EB) [92], MOPED [93] used deterministic weights, which were derived from a pretrained DNN with the same architecture, to select meaningful prior distributions over the weight space. Later, in [94], an approach based on parametric EB was integrated into MOPED for mean-field VI in Bayesian DNNs (BDNNs), and it used a fully factorized Gaussian distribution to model the weights. In addition, a real-world case study,

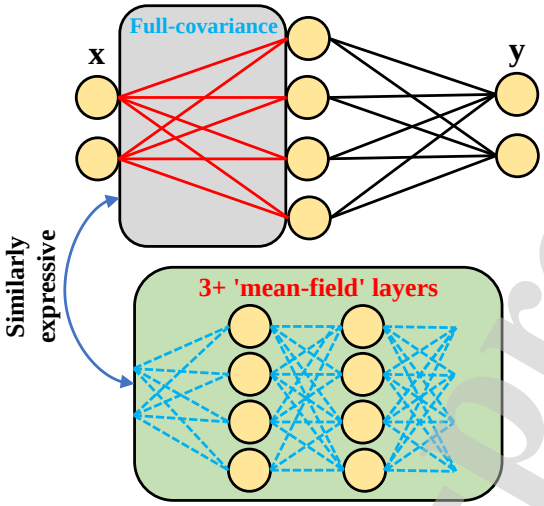


Fig. 8: The general architecture of the deeper linear mean-field network with three or more mean-field weight layers, which is reproduced based on [103].

i.e., diabetic retinopathy diagnosis, was used to evaluate the method. Subedar et al. [95] proposed an uncertainty-aware framework based on multimodal Bayesian fusion for activity recognition. They scaled a BDNN to a deeper structure by combining deterministic and variational layers. Marino et al. [96] proposed a stochastic modeling-based approach to model uncertainty. Specifically, a deep BNN (DBNN) was used to perform stochastic learning for the system. A variational BNN [97], which is a generative-based model, was proposed to predict the superconducting transition temperature. Specifically, VI was adapted to compute the distribution in the latent space for the model.

Louizos and Welling [98] adopted stochastic gradient VI [99] to compute the posterior distributions over the weights of NNs. Hubin and Storvik [100] proposed a stochastic VI method that jointly considers both model and parameter uncertainties in BNNs, and they introduced latent binary variables to include or exclude certain weights of the model. Liu et al. [101] integrated VI into a spatial–temporal NN to approximate the posterior parameter distribution of the network and estimate the probability of the prediction. Ryu et al. [102] integrated a graph convolutional network (GCN) into the Bayesian framework to learn representations and predict molecular properties. Swiatkowski et al. [89] empirically studied Gaussian mean-field VI. They decomposed the variational parameters into a low-rank factorization to obtain a more compact approximation and improved the signal-to-noise ratio (SNR) of the stochastic gradient (SG) in estimating the lower bound of the variation. Franquhar et al. [103] used mean-field VI to better train deep models. They argued that a deeper linear mean-field network can provide an analogous distribution of the function space to that of shallowly full-covariance networks. A schematic view of the proposed approach is shown in Fig. 8.

3.4. Bayesian Active Learning (BAL)

AL methods aim to learn from unlabeled samples by querying an oracle [104]. Defining the right acquisition function, i.e., the condition under which a sample is most informative for the model, is the main challenge for AL-based methods. Although existing AL frameworks have shown promising results in a variety of tasks, they lack scalability to high-dimensional data [105]. In this regard, Bayesian approaches can be integrated into the DL structure to represent uncertainty and then combined with a deep AL acquisition function to probe for uncertain samples in the oracle.

Deep Bayesian AL (DBAL) [106] combines an AL framework with Bayesian DL to deal with high-dimensional data problems, i.e., image data. DBAL uses batch acquisition to select the n samples with the highest Bayesian AL by disagreement (BALD) [107] score. Model priors from empirical Bayes (MOPED) [108] uses BALD to evaluate uncertainty. In addition, MC dropout has been applied to estimate model uncertainty. Later, Krisch et al. [109] proposed BatchBALD, which uses a greedy algorithm to select a batch in linear

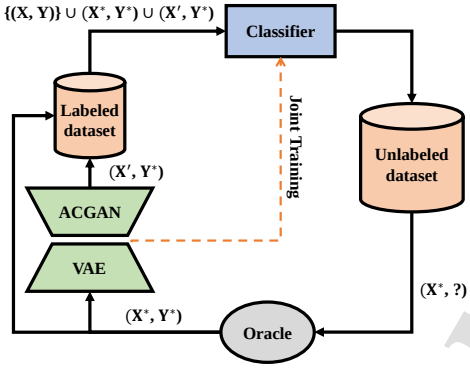


Fig. 9: Bayesian generative active deep learning (ACGAN stands for the auxiliary-classifier GAN), which is reproduced based on [121].

time, reducing the runtime. They modeled uncertainty by leveraging Bayesian AL (BAL) using dropout sampling. In [110], two types of uncertainty measures, namely, entropy and BALD [107], were compared. ActiveHARNet [111], which is an AL-based framework for human action recognition, modeled uncertainty by linking BNNs with GP using dropout. To achieve this, dropout was applied before each fully connected layer to estimate the mean and variance of the BNN. DeepBASS [112], i.e., a deep AL semisupervised learning method, is an expectation-maximization-based [113] technique paired with an AL component. It applies MC dropout to estimate uncertainty.

Scandalea et al. [114] proposed a framework based on the U-Net structure for deep AL to segment biomedical images and used an uncertainty measure obtained by MC dropout to suggest samples to be annotated. Specifically, uncertainty was defined based on the standard deviation (STD) of the posterior probabilities of the MC samples. Zheng et al. [115] varied the number of Bayesian layers and their positions to estimate uncertainty through AL on the MNIST dataset. The outcome indicated that a few Bayesian layers near the output layer are enough to fully estimate the uncertainty of the model.

Inspired by [116], Bayesian batch AL [117], which selects a batch of samples at each AL iteration to perform posterior inference over the model parameters, was proposed for large-scale problems. Active user training [118], which is a BAL-based crowdsourcing model, was proposed to address high-dimensional and complex classification problems. In addition, the Bayesian inference method proposed in [119] was used to consider the uncertainty of the confusion matrix of the annotators.

Several generative-based AL frameworks have been introduced. In [120], a semi-supervised Bayesian AL model, which is a deep generative-based model that uses BNNs to obtain the discriminative component, was developed. Tran et al. [121] proposed Bayesian-based generative deep AL (BGADL) (Fig. 9) for image classification problems. They first used the concept of DBAL to select the most informative samples, and then VAE-ACGAN was applied to generate new samples based on the selected ones. Akbari et al. [122] proposed a unified BDL framework to quantify both aleatoric and epistemic uncertainty for activity recognition. They used an unsupervised DL model to extract features from the time series, and then their posterior distributions were learned through a VAE model. Finally, dropout [35] was applied after each dense layer and test phase for the randomness of the model weights and samples from the approximate posterior, respectively.

3.5. Bayes by Backprop (BBB)

The process of learning a probability distribution using the weights of NNs plays a significant role in obtaining better prediction results. Blundell et al. [123] proposed a novel yet efficient algorithm named Bayes by Backprop (BBB) to quantify the uncertainty of these weights. The proposed BBB minimizes the compression cost, which is known as the variational free energy (VFE) or the (expected) lower bound of the marginal likelihood. To do this, they defined a cost function as follows:

$$F(\mathcal{D},\theta) = \mathrm{KL}[q(\mathbf{w}|\theta)\|P(\mathbf{w})] - \mathbb{E}_{q(\mathbf{w},\theta)}[\log P(\mathcal{D}|\mathbf{w})]. \tag{21}$$

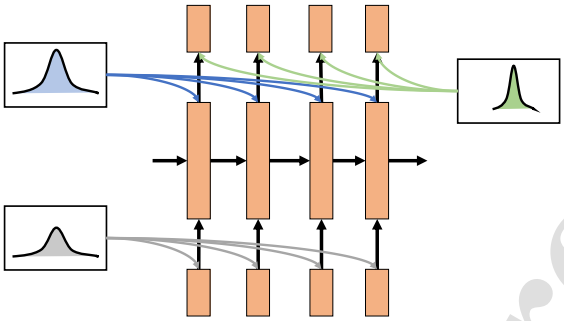


Fig. 10: Bayesian RNNs (BRNNs); the image is reproduced based on the model proposed by Fortunato et al. [124].

The BBB algorithm uses the unbiased gradient estimates of the cost function in (21) to learn a distribution over the weights of NNs. In another work, Fortunato et al. [124] proposed a new Bayesian RNN (BRNN) using the BBB algorithm. To improve the BBB algorithm, they used a simple adaptation of truncated backpropagation throughout time. The proposed BRNN model is shown in Fig. 10.

Ebrahimi et al. [125] proposed an uncertainty-guided continual approach with BNNs (called the uncertainty-guided continual learning (CL) technique with BNNs (UCB)). CL enables learning a variety of new tasks while incorporating previously learned ones. The proposed UCB exploits the predicted uncertainty of the posterior distribution to formulate the modification of "important" parameters both by setting a hard threshold as well as in a soft way. Recognizing different actions in videos not only requires big data but is also a time-consuming process. To address this issue, de la Riva and Mettes [126] proposed a Bayesian-based deep learning method (Bayesian 3D ConvNet) to analyze a small number of videos. In this regard, BBB was extended for use by 3D CNNs and then employed to deal with uncertainty over the convolution weights in the proposed model. To this end, a Gaussian distribution was applied to approximate the correct posterior in the proposed 3D convolution layers using the mean and STD as follows:

$$\begin{cases} \theta = (\mu, \alpha), \\ \sigma^2 = \alpha.\mu^2, \\ q_\theta(\mathbf{w}_{ijhwt}|\mathcal{D}) = \mathcal{N}(\mu_{ijhwt}, \alpha_{ijhwt}\mu^2_{ijhwt}), \end{cases} \tag{22}$$

where i represents the input, j is the output, h is the filter height, w is the filter width and t is the time dimension. In another work, Ng et al. [66] compared the performance of two well-known uncertainty methods (MC dropout and BBB) in medical image segmentation (cardiac MRI) on a U-Net model. The results showed that MC dropout and BBB had similar performances in medical image segmentation tasks.

3.6. Variational Autoencoders

An autoencoder is a variant of DL that consists of two components: (i) an encoder and (ii) a decoder. The encoder aims to map high-dimensional input samples x to a low-dimensional latent variable z, while the decoder reproduces the original samples x using the latent variable z. The latent variables are compelled to conform a given prior distribution $P(z)$. Variational autoencoders (VAEs) [99] are effective methods to model the posterior. They cast learning representations for high-dimensional distributions as a VI problem [128]. A probabilistic model $P_\theta(x)$ of sample x in a data space with a latent variable z in a latent space can be written as follows:

$$p_\theta(x) = \int_z p_\theta(x|z)p(z). \tag{23}$$

VI can be used to model the evidence lower bound $\log p_\theta(x)$ as follows:

$$\log p_\theta(x) = \mathbb{E}_{q_\phi(z|x)}[\log p_\theta(x|z)] - D_{KL}(q_\phi(z|x)\|p(x)), \tag{24}$$

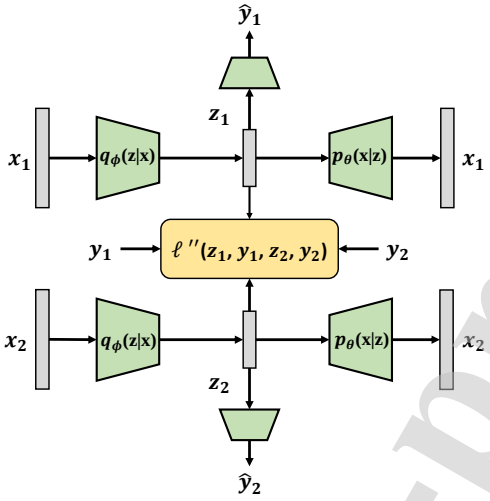


Fig. 11: Pairwise supervised hashing-Bernoulli VAE (PSHBVAE), which is adopted from [127].

where $q_\phi(z|x)$ and $p_\theta(x|z)$ are the encoder and decoder models, respectively, and ϕ and θ indicate their parameters.

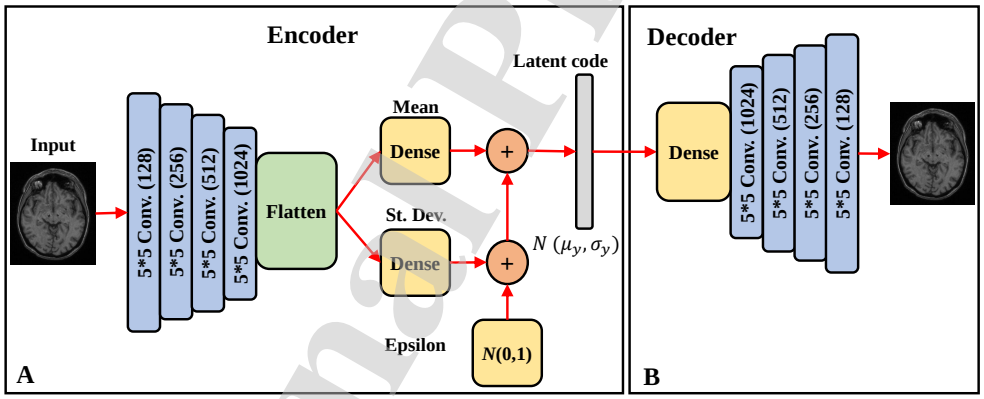
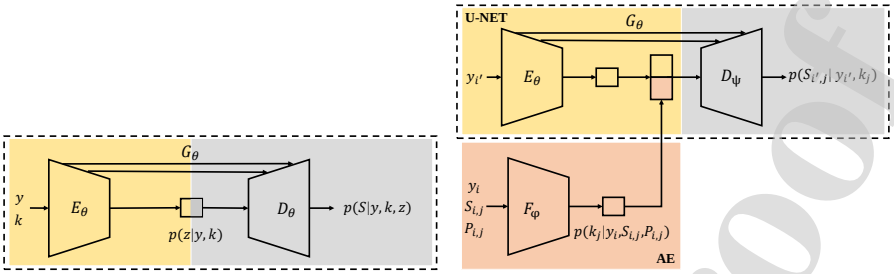


Fig. 12: A schematic view of the VAE model proposed by Edupuganti et al., which is reproduced based on [129].

Zamani et al. [127] developed a discrete VAE framework with Bernoulli latent variables as binary hashing code (Fig.11). The stochastic gradient was exploited to learn the model. They proposed a pairwise supervised hashing (PSH) framework to derive better hashing codes. PSH maximizes the ELBO with weighted KL regularization to learn more informative binary codes, and it adapts a pairwise loss function to reward within-class similarity and between-class dissimilarity to minimize the distances among the hashing codes of samples from the same class and vice versa.

Bohm et al. [130] studied UQ for linear inverse problems using VAEs. Specifically, a vanilla VAE with a mean-field Gaussian posterior was trained on uncorrupted samples under the ELBO. In addition, the EL_2O method [131] was adopted to approximate the posterior. Edupuganti et al. [129] studied UQ tasks in magnetic resonance image recovery (see Fig. 12). To this end, a VAE-GAN, which is a probabilistic recovery scheme, was developed to map low-quality images to high-quality ones. The VAE-GAN consisted of a VAE and a multilayer CNN as the generator and discriminator, respectively. In addition, Stein's unbiased risk estimator (SURE) was leveraged as a proxy to predict error and estimate the uncertainty of the model.



(a) U-Net (b) VUNet

Fig. 13: A general view of U-Net and VUNet which are reproduced based on [132].

In [132], a framework based on variational U-Net [133] architecture was proposed for UQ tasks in reservoir simulations. Both a simple U-Net and variational U-Net (VUNet) are illustrated in Fig. 13. Cosmo VAE [134], which is a VAE based on DL, i.e., U-Net, was proposed to restore missing observations in the cosmic microwave background (CMB) map. As such, variational Bayes approximation was used to determine the ELBO of the likelihood of the reconstructed image. Mehrasa et al. [135] proposed the action point process VAE (APP VAE) for action sequences. APP VAE consists of two LSTMs to estimate the prior and posterior distributions. Sato et al. [136] proposed a VAE-based UA for anomaly detection. They used MC sampling to estimate the posterior.

Since VAEs are not stochastic processes, they are limited to encoding finite-dimensional priors. To address this limitation, Mishra et al. [137] developed the prior encoding VAE, i.e., πVAE. Inspired by the Gaussian process [138], πVAE is a stochastic process that learns a distribution over functions. To achieve this, the πVAE encoder first transforms the locations to a high-dimensional space and then uses a linear mapping to link the feature space to the outputs. The πVAE encoder aims to recreate a linear mapping from a lower-dimensional probabilistic embedding. Finally, the recreated mapping is used to obtain the reconstruction of the outputs. Guo et al. [139] used a VAE to deal with data uncertainty under a just-in-time learning framework. A Gaussian distribution was employed to describe the latent space features variablewise, and then KL divergence was used to ensure that the selected samples were the most relevant to a new sample. Daxberger et al. [140] tried to detect OoD samples during the test phase. To this end, an unsupervised, probabilistic framework was developed based on a Bayesian VAE. In addition, they estimated the posterior over the decoder parameters by applying SG-MCMC.

3.7. Laplacian Approximations

Laplacian approximations (LAs) are popular UQ methods that are used to estimate Bayesian inferences [141]. They build a Gaussian distribution around the true posterior using a Taylor expansion around the MAP, θ^*, as follows:

$$p(\theta|D) \approx p(\theta^*)\exp\{-\frac{1}{2}(\theta-\theta^*)'\mathbb{H}_{|\theta^*}(\theta-\theta^*)\}, \tag{25}$$

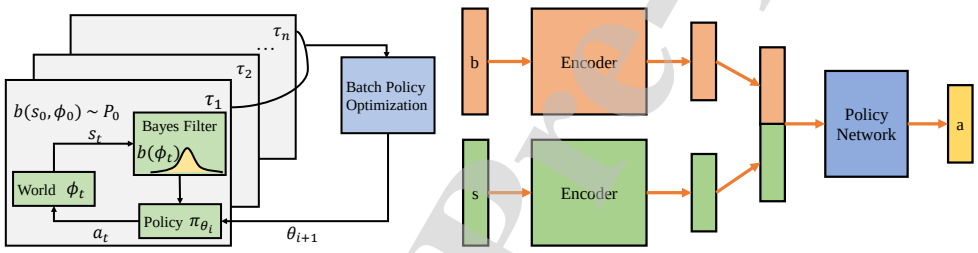
where $\mathbb{H}_{|\theta} = \nabla_\theta p(y|\theta)\nabla_\theta p(y|\theta)'$ indicates the Hessian of the likelihood estimated at the MAP estimate. Ritter et al. [142] introduced a scalable LA (SLA) approach for different NNs. They proposed the model, then compared it with other well-known methods such as dropout and a diagonal LA for uncertainty estimation in networks. Additional studies have been performed on LA, such as [143, 144, 145, 146]. For example, Shinde et al. [143], with the help of conditional random fields (CRFs) on top of BNNs, could determine contextual information and carry out semisupervised learning. Then, the authors compared the performance of LA with a variant of MC dropout. Shinde et al., in another study [144], evaluated the performance of LA on an autonomous driving application (the KITTI dataset was used). Their findings showed that uncertainty can be meaningful; however, they encouraged conducting more experimental evaluations on this point. In another study, Lee et al. [145] used an LA-based inference engine for natural parameters and information

in the form of a Gaussian distribution. The authors managed to scale LA on the ImageNet dataset by spending considerable time tuning the hyperparameters so that they could make a meaningful comparison. Finally, Humt et al. [146] applied existing BO techniques to tune the hyperparameters of LA. The outcomes indicated that the proposed BO approach required fewer iterations than random search.

3.8. Uncertainty Quantification in Reinforcement Learning

Uncertainty plays a key role in estimating the method's performance in various fields (e.g. in RL [147]). Different UQ methods in RL have been widely-investigated in the literature [148]. Lee et al. [149] formulated the model uncertainty problem as a Bayes-adaptive Markov decision process (BAMDP). The general BAMDP is defined by a tuple $\langle \mathrm{S}, \Phi, \mathrm{A}, \mathrm{T}, \mathrm{R}, P_0, \gamma \rangle$, where S represents the underlying MDP's observable state space, Φ indicates the latent space, A represents the action space, T is the parameterized transition and R is the reward function. Let b_0 be an initial belief; a Bayes filter updates the posterior as follows:

$$b'(\phi'|s, b, a', s') = \eta \sum_{\phi \in \Phi} b(\phi) T(s, \phi, a', s', \phi') \tag{26}$$



(a) Training procedure (b) Network structure

Fig. 14: A general view of BPO which is reproduced based on [149].

Then, the Bayesian policy optimization (BPO) method (see Fig. 14) is applied to the POMDPs as a Bayes filter to compute the belief b of the hidden state as follows:

$$b'(s') = \psi(b, a', o') = \eta \sum_{s \in S} b(s) T(s, a', s') Z(s, a', o') \tag{27}$$

In another work, O'Donoghue et al. [150] proposed the uncertainty Bellman equation (UBE) to quantify uncertainty. The authors used a Bellman-based equation that propagated the uncertainty (here, variance) relationship of the posterior distribution of the Bayesian model. Kahn et al. [151] presented a new UA model for a learning algorithm to control a mobile robot. A review of past studies in RL shows that different Bayesian approaches have been used for handling parameter uncertainty [152]. Bayesian RL was thoroughly reviewed by Ghavamzadeh et al. [152] in 2015. Due to page limitation, we do not discuss here the application of UQ to RL, but we summarize some of the recent studies here.

Kahn et al. [151] used both bootstrapping and dropout methods to estimate uncertainty in NNs and then used them in a UA collision prediction model. In addition to Bayesian statistical methods, ensemble methods have been used to quantify uncertainty in RL [153]. In this regard, Tschantz et al. [153] applied an ensemble of different point-estimate parameters $\theta = \{\theta_0, ..., \theta_B\}$ that were trained on various batches of a dataset D and then maintained and treated by the posterior distribution $p(\theta|D)$. The ensemble method helped to capture both aleatoric and epistemic uncertainty. There are other UQ techniques used in RL; however, we are not able to discuss all of them in detail in this work for various reasons, such as page restrictions and the breadth of the article. Table 2 summarizes different UQ methods used in a variety of RL subjects.

Table 2: Further information on some UQ methods used in RL.

Study	Year	Application	Goal/Objective	UQ method	Code
Kalweit and Boedecker [154]	2017	Continuous Deep RL (CDRL)	Minimizing real-world interaction	Model-assisted Bootstrapped Deep Deterministic Policy Gradient (MA-BDDPG)	×
Tegho et al. [155]	2018	Dialogue management context	Dialogue policy optimisation	BBB propagation deep Q-networks (BBQN)	×
Riquelme et al. [156]	2018	Approximating the posterior sampling	Balancing both exploration and exploitation in different complex domains	Deep Bayesian Bandits Showdown using Thompson sampling	✓
Riquelme et al. [156]	2018	Approximating the posterior sampling	Balancing both exploration and exploitation in different complex domains	Deep Bayesian Bandits Showdown using Thompson sampling	✓
Janz et al. [157]	2019	Temporal difference learning	Posterior sampling for RL (PSRL)	Successor Uncertainties (SU)	✓
Shen and How [158]	2019	Discriminating potential threats	Stochastic belief space policy	Soft-Q learning	×
Benatan and Pyzer-Knapp [159]	2019	Safe RL (SRL)	The weights in RNN using mean and variance weights	Probabilistic Backpropagation (PBP)	×
Huang et al. [160]	2019	Model-based RL (MRL)	Better decision and improve performance	Bootstrapped model-based RL (BMRL)	×
Eriksson and Dimitrakakis [161]	2019	Risk measures and leveraging preferences	Risk-Sensitive RL (RSRL)	Epistemic Risk Sensitive Policy Gradient (EPPG)	×
Lötjens et al. [162]	2019	SRL	UA navigation	Ensemble of MC dropout (EMCD) and Bootstrapping	×
Clements et al. [163]	2019	Designing risk-sensitive algorithm	Disentangling aleatoric and epistemic uncertainties	Combination of distributional RL (DRL) and Approximate Bayesian computation (ABC) methods with NNs	✓
Metelli et al. [164]	2019	Temporal difference learning	Balancing exploration and exploitation	WQL: Wasserstein Q-Learning	✓
D'Eramo et al. [165]	2019	Drive exploration	Multi-Armed Bandit (MAB)	Bootstrapped deep Q-network with TS (BDQNTS)	×
Tschantz et al. [166]	2020	Model-based RL	Exploration and exploitation	Scaling active inference	×
Lee et al. [167]	2020	Multi-agent RL	Lack of entirely accurate knowledge of each agent in the model	robust Markov game	×
Ramakrishnan et al. [168]	2020	Blind spot detection	Mismatch between training and execution environments	Safe sim-to-real transfer	×

4. Ensemble Techniques

Deep NNs (DNNs) have been effectively employed in a wide variety of machine learning tasks and have achieved state-of-the-art (SOTA) performance in different domains such as bioinformatics, NLP, speech recognition and computer vision [169, 170]. In supervised learning benchmarks, NNs yielded competitive accuracy but poor predictive uncertainty quantification. Hence, they are inclined to generate overconfident predictions. Incorrect overconfident predictions can be harmful; therefore, it is important to handle UQ in a proper manner in real-world applications [171]. As empirical evidence of uncertainty estimates are not available in general, the quality of predictive uncertainty evaluation is a challenging task. Two evaluation measures, calibration and domain shift, are applied, which are usually inspired by the practical applications of NNs. Calibration measures the discrepancy between long-run frequencies and subjective forecasts. The second notion concerns the generalization of predictive uncertainty to a domain shift that estimates whether the network knows what it knows. An ensemble of models enhances the predictive performance. However, it is not evident why and when an ensemble of NNs can generate good uncertainty estimates. Bayesian model averaging (BMA) believes that the true model lies within the hypothesis class of the prior and executes soft model selection to locate the single best model within the hypothesis class. In contrast, ensembles combine models to discover more powerful models; ensembles can be anticipated to be better when the true model does not lie within the hypothesis class.

The authors of [172] devised the maximize overall diversity (MOD) model to estimate ensemble-based uncertainty by taking into account the diversity in ensemble predictions across possible future inputs. Gustafsson et al. [173] presented an evaluation approach for measuring uncertainty estimations to investigate robustness in the computer vision domain.

Chua et al. [174] devised a novel method called probabilistic ensembles with trajectory sampling that integrated sampling-based uncertainty propagation with a UA deep network dynamics approach. The au-

thors of [169] demonstrated that prevailing calibration error estimators were unreliable in the small data regime and hence proposed a kernel density-based estimator for calibration performance evaluation and proved its consistency and unbiasedness. Liu et al. [175] presented a Bayesian nonparametric ensemble method that enhanced an ensemble model that augmented a model's distribution functions using Bayesian nonparametric machinery and a prediction mechanism. In another study, the researchers [176] exploited the challenges associated with attaining uncertainty estimations for structured predictions and presented baselines for sequence-level out-of-domain input detection, sequence-level prediction rejection and token-level error detection utilizing ensembles.

Ensembles involve memory and computational cost, which is not acceptable in many applications [177]. There has been noteworthy work done on the distillation of an ensemble into a single model. Such approaches achieved comparable accuracy using ensembles and mitigated the computational costs. In the posterior distribution $p(\boldsymbol{\theta}|D)$, the uncertainty of a model is captured. Let us consider the posterior of a sampled ensemble of models $\{P(y|x^\star, \boldsymbol{\theta}^{(m)})\}_{m=1}^M$ as follows [177]:

$$\{P(y|x^\star, \boldsymbol{\theta}^{(m)})\}_{m=1}^M \rightarrow \{P(y|\pi^{(m)})\}_{m=1}^M, \pi^m = f(x^\star; \theta^{(m)}), \boldsymbol{\theta}^{(m)} \sim p(\boldsymbol{\theta}|D), \tag{28}$$

where a test value x^* is input and π represents the parameters of a categorical distribution $[P(y = w_1), ..., P(y = w_k)]^T$. By taking into account the expectation with respect to the model posterior, predictive posterior or expected predictive distribution, a test input x^* is acquired. Then we have:

$$P(y|x^\star, D) = \mathbb{E}_{p(\boldsymbol{\theta}|D)}[P(y|x^\star, \boldsymbol{\theta})]. \tag{29}$$

Different estimates of data uncertainty are demonstrated by each of the models $P(y|x^\star, \boldsymbol{\theta}^{(m)})$. The 'disagreement' or the level of spread of an ensemble sampled from the posterior is due to the uncertainty in the predictions as a result of model uncertainty. Let us consider an ensemble $\{P(y|x^\star, \boldsymbol{\theta}^{(m)})\}_{m=1}^M$ that yields the expected set of behaviors; the entropy of the expected distribution $P(y|x^\star, D)$ can be utilized as an estimate of the total uncertainty in the prediction. Measures of the spread or 'disagreement' of the ensemble such as MI can be used to assess the uncertainty in the predictions due to knowledge uncertainty as follows:

$$\underbrace{\mathcal{M}I[y, \boldsymbol{\theta}|x^\star, D]}_{Knowledge\ Uncertainty} = \underbrace{H[\mathbb{E}_{p(\boldsymbol{\theta}|\mathcal{D})}[P(y|x^\star, \boldsymbol{\theta})]]}_{Total\ Uncertainty} - \underbrace{\mathbb{E}_{p(\boldsymbol{\theta}|\mathcal{D})}[H[P(y|x^\star, \boldsymbol{\theta})]]}_{Expected\ Data\ Uncertainty}. \tag{30}$$

The total uncertainty can be decomposed into the expected data uncertainty and knowledge uncertainty via MI formulation. If the model is uncertain, both in out-of-domain areas and regions of severe class overlap, the entropy of the total uncertainty or predictive posterior is high. If the models disagree, the difference of the expected entropy and the entropy of the predictive posterior of the individual models will be nonzero. For example, the MI will be low and the expected and predictive posterior entropy will be similar, and each member of the ensemble will demonstrate a high-entropy distribution in regions of class overlap. In such scenarios, data uncertainty dominates total uncertainty. The predictive posterior is nearly uniform while the expected entropy of each model may be low; this is the result of diverse distributions over classes as a result of out-of-domain inputs. In this region of input space, knowledge uncertainty is high because the model's understanding of the data is low. In ensemble distribution distillation, the aim is to capture not only the diversity but also the mean of the ensemble. An ensemble can be observed as a set of samples from an implicit distribution of the output distributions:

$$\{P(y|x^\star, \boldsymbol{\theta}^{(m)})\}_{m=1}^M \rightarrow \{P(y|\pi^{(m)})\}_{m=1}^M, \pi^{(m)} \sim p(\pi|x^\star, \mathcal{D}). \tag{31}$$

Prior networks, a new class model, were proposed to explicitly parameterize a conditional distribution over output distributions $p(\pi|x^\star, \hat{\phi})$ utilizing a single NN parameterized by a point estimate of the model parameters $\hat{\phi}$. An ensemble can be emulated effectively by a prior network and hence show the same measure

of uncertainty. By parameterizing the Dirichlet distribution, the prior network $p(\pi|x^{\star}, \hat{\phi})$ represents a distribution over categorical output distributions. The ensembling performance is measured by an uncertainty estimation. DL ensembles produce benchmark results in uncertainty estimation.

Different sources of model uncertainty can be addressed by incorporating one of the presented ensemble techniques to obtain the Bayesian nonparametric ensemble (BNE) model devised by Liu et al. [175]. Bayesian nonparametric machinery was utilized to augment distribution functions and predict a model with a BNE. The BNE measures the uncertainty patterns in the data distribution and decomposes the uncertainty into discrete components that are due to error and noise. The model yielded precise uncertainty estimates from observational noise and demonstrated its utility with respect to model bias detection and uncertainty decomposition for an ensemble method used in prediction. The predictive mean of a BNE can be expressed as follows [175]:

$$E(y|X,\omega,\delta,G) = \sum_{k=1}^{K} f_k(X)\omega_k + \underbrace{\delta(X)}_{Due\ to\ \delta} + \underbrace{\int_{y\in\dagger}\Big[\Phi((y|X,\mu) - G[\Phi((y|X,\mu)\Big]dy}_{Due\ to\ G}. \tag{32}$$

The predictive mean for the full BNE is composed of three sections:

1. The predictive mean of the original ensemble is $\sum_{k=1}^{K} f_k(X)\omega_k$;
2. The BNE's direct correction to the prediction function is represented by the term δ; and
3. The BNE's indirect correction to the prediction derived from the relaxation of the Gaussian assumption in the model cumulative distribution function is represented by the term $\int\Big[\Phi((y|X,\mu) - G[\Phi((y|X,\mu)\Big]dy$. In addition, two error correction terms $\mathcal{D}_{\delta}(y|X)$ and $\mathcal{D}_{G}(y|X)$ are presented.

To denote the BNE's predictive uncertainty estimation, the term $\Phi_{\varepsilon,\omega}$ is used, which is the predictive cumulative distribution function of the original ensemble (i.e., including the variance σ_{ε}^2 and mean $\sum_k f_k\omega_k$). The BNE's predictive interval is represented as [175]:

$$U_q(y|X,\omega,\delta,G) = \left[\Phi_{\varepsilon,\omega}^{-1}\left(G^{-1}(1-\frac{q}{2}|X)\right) + \delta(x), \Phi_{\varepsilon,\omega}^{-1}\left(G^{-1}(1+\frac{q}{2}|X)\right) + \delta(x)\right]. \tag{33}$$

Comparing the above equation to the predictive interval of the original ensemble $\left[\Phi_{\varepsilon,\omega}^{-1}\left(G^{-1}(1-\frac{q}{2}|X)\right), \Phi_{\varepsilon,\omega}^{-1}\left(G^{-1}(1+\frac{q}{2}|X)\right)\right]$, it can be observed that the residual process δ adjusts the locations of the BNE predictive interval endpoints while G calibrates the spread of the predictive interval.

Finally, Wen et al. [178] designed an ensemble method, called BatchEnsemble, whose memory and computational costs are considerably lower than those of typical ensembles. BatchEnsemble achieves this by denoting each weight matrix as the Hadamard product of a shared weight among all ensemble members and considering a rank-one matrix for each ensemble member. BatchEnsemble is not only parallelizable across devices, where one device trains one ensemble member, but also parallelizable within a device, where multiple ensemble members are updated concurrently for a given mini-batch.

As an important part of ensemble techniques, loss functions play a significant role of having a good performance by different ensemble techniques. In other words, choosing the appropriate loss function can dramatically improve results. We summarise the most important loss functions applied for UQ in Table A.1 in the Appendix.

4.1. Deep Ensemble

Deep ensemble (DE) is another powerful method for measuring uncertainty that has been used extensively in many real-world applications [179]. To achieve good learning results, the data distributions in the testing datasets should be as close as in the training datasets. In many situations, the distributions of the test

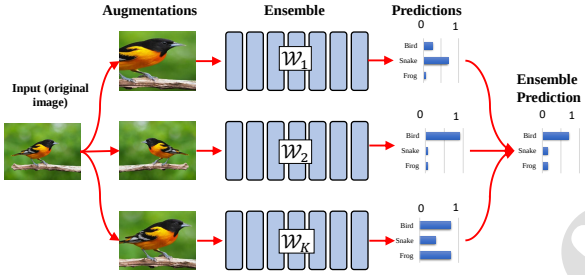


Fig. 15: A schematic view of TTA for ensembling techniques, which is reproduced based on [181].

datasets are unknown, especially in cases of uncertainty prediction problems. Hence, it is difficult for traditional learning models to yield competitive performance. Some researchers have applied MCMC and BNNs that relied on the prior distributions of datasets to solve uncertainty prediction problems [172]. When these approaches are employed in large networks, it becomes computationally expensive. Model ensembling is an effective technique that can be used to enhance the predictive performance of supervised learners. DEs are applied to obtain better predictions on the test data and produce model uncertainty estimates when the learners are provided with OoD data. The success of the ensembles depends on the variance reduction generated by combining predictions that are prone to several types of errors individually. Hence, the improvement in the predictions is achieved by utilizing a large ensemble with numerous base models, and such ensembles also generate distributional estimates of model uncertainty.

McDermott and Wikle [180] proposed a DE echo state network model for spatiotemporal forecasting in uncertainty quantification. Hu et al. [179] proposed a model called margin-based Pareto DE pruning utilizing a DE network that yielded competitive uncertainty estimations with elevated confidence of the prediction interval coverage probability and a small value of the prediction interval width. The authors of [181] exploited in-domain uncertainty, examined the standards for its quantification and revealed the pitfalls of prevailing matrices. They presented the DE equivalent score (DEE) and demonstrated how an ensemble of only a few trained networks can be equivalent to many refined ensembling methods with respect to test performance. For one ensemble, they proposed test-time augmentation (TTA) to improve the performance of different ensemble learning techniques (see Fig. 15). There have been other studies on TTA for quantifying uncertainties, such as [182, 38, 183, 184].

Wilson and Izmailov [185] showed how DE models can provide a better approximation of the Bayesian model average in DL than standard Bayesian methods (such as BBB). In other words, they believed that DEs can be considered a Bayesian approach. Moreover, the authors discussed the importance of multi-basin marginalization (by applying a new procedure called MultiSWAG) for epistemic uncertainty representation, robustness to data corruption and ultimately performance. They also showed that it can entirely eliminate double descent behavior. They analyzed priors over functions $P(f)$ when induced by Gaussian priors over different parameters of NNs. Based on this procedure, they could show that the proposed process has several significant properties, such as the following:

1. A valuable induced correlation function for a wide variety of images;
2. An impressive ability to generate beneficial features;
3. Excellent support for structured datasets versus corrupted datasets through marginal likelihood;
4. A significant performance improvement over standard training.

The obtained outcomes showed how probability UQ in DL can help provide a prescription for model construction that eliminates the ambiguous generalization properties of DNNs, such as having the capacity to fit images with random classes (labels), double descent, and overparametrization.

However, DEs [186] are a simple approach that presents independent samples in various modes of loss setting. Under a fixed test-time computed budget, DEs can be regarded as a powerful baseline for judging the performance of other ensembling methods. It is a challenging task to compare the performance of ensembling methods. Different values of matrices are achieved by different models on different datasets. Interpretability is lacking for the values of matrices, as performance gains are compared with dataset- and model-specific

baselines. Hence, Ashukha et al. [181] proposed DDE with the aim of introducing interpretability and a perspective that applies DEs to compute the performance of other ensembling methods. The DDE score aims to answer the question: what size of DE demonstrates the same performance as a specific ensembling technique? The DDE score is based on calibrated log-likelihood (CLL). DDE is defined for an ensembling technique (m), and the lower and upper bounds are depicted as follows [181]:

$$\mathcal{DEE}_m(k) = min\{l \in R, l \geq 1 | CLL_{DE}^{mean}(l) \geq CLL_m^{mean}(k)\}, \tag{34}$$

$$\mathcal{DEE}_m^{upper/lower}(k) = min\{l \in R, l \geq 1 | CLL_{DE}^{mean}(l) \mp CLL_{DE}^{std}(l) \geq CLL_m^{mean}(k)\}, \tag{35}$$

where the mean and standard deviation of the calibrated log-likelihood yielded by an ensembling technique m with l samples is denoted as $CLL_m^{mean/std}(l)$. They measured $CLL_{DE}^{std}(l)$ and $CLL_{DE}^{mean}(l)$ for natural numbers $l \in \mathbb{N}_{>0}$, and linear interpolation was applied to define them for real values $l \geq 1$. They depicted $DDE_m(k)$ for different numbers of samples k for different methods m with upper and lower bounds $DEE_m^{upper}(k)$ and $DEE_m^{lower}(k)$.

A DE echo state network (D-EESN) model with two versions for spatiotemporal forecasting and the associated uncertainty measurement was presented in [180]. The first framework applies a bootstrap ensemble approach, and the second was devised within a hierarchical Bayesian framework. Multiple levels of uncertainties and non-Gaussian data types were accommodated by the general hierarchical Bayesian approach. The authors of [180] broadened some of the deep ESN technique constituents presented by Antonelo et al. [187] and Ma et al. [188] to fit within a spatiotemporal ensemble approach in the D-EESN model to contain this kind of structure. As in the previous section, we summarize a few loss functions of DEs in Table A.2 in the Appendix.

4.2. Deep Ensemble Bayesian/Bayesian Deep Ensemble

The expressive power of various ensemble techniques such as DE Bayesian (DEB)/Bayesian DE (BDE) is shown extensively in the literature. However, traditional learning techniques suffered from several drawbacks and limitations, as listed in [189]. To overcome these limitations, Fersini et al. [189] utilized the ensemble learning approach to mitigate the noise sensitivity related to language ambiguity, and more accurate predictions of polarity could be estimated. The proposed ensemble method employed Bayesian model averaging, where both the reliability and uncertainty of each single model were considered. The study [190] presented one alteration to the prevailing approximate Bayesian inference method by regularizing the parameters related to values derived from a distribution that could be set equal to the prior. The analysis of the process suggested that the recovered posterior was centered correctly but leaned toward an overestimated correlation and underestimated marginal variance. To obtain uncertainty estimates, one of the most promising frameworks is Deep BAL (DBAL) with MC dropout.

Pop et al. [186] argued that in VI methods, the mode collapse phenomenon was responsible for the overconfident predictions of DBAL methods. They devised DE-BAL, which addressed the mode collapse issue and improved the MC dropout method. In another study, Pop et al. [191] proposed a novel AL technique especially for DNNs. The statistical properties and expressive power of model ensembles were employed to enhance the SOTA deep BAL technique that suffered from the mode collapse problem. In another work, Pearce et al. [192] proposed a new ensemble of NNs, an approximate Bayesian ensembling approach called "*anchoredensembling*". The proposed approach regularized the parameters regarding values attracted from a distribution.

4.3. Uncertainty in Dirichlet Deep Networks

Tsiligkaridis [193] devised a new technique, information-aware Dirichlet networks, that learned an explicit Dirichlet prior distribution on predictive distributions by lessening a bound on the probable maximum norm of the prediction error and penalizing information related to incorrect results. Enhanced uncertainty estimation was achieved by deriving the properties of the new cost function. To detect adversarial examples and estimate out-of-distribution and within-distribution uncertainty, the method of Tsiligkaridis [193] showed better performance than many SOTA NNs. The standard approach is to train deterministic neural nets to

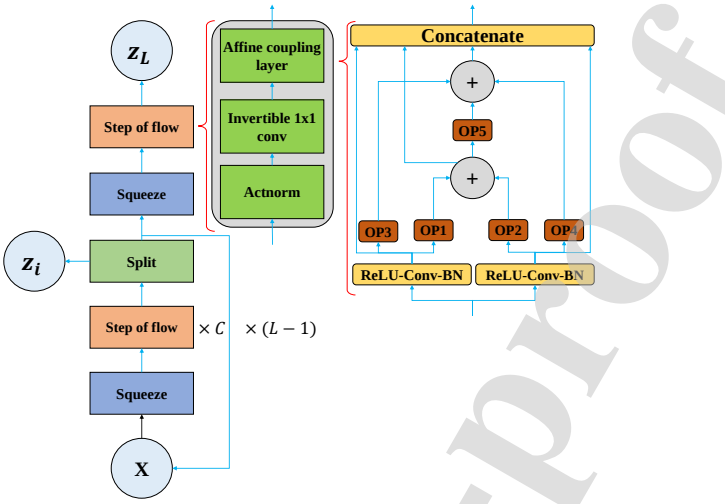


Fig. 16: A single block diagram for searching the space in the architecture, which is reproduced based on [195].

reduce prediction loss. The resulting model remains ignorant to the prediction confidence. Orthogonally to BNNs, which indirectly infer the uncertainty of predictions through weight uncertainties, Sensoy et al. [194] presented explicit modeling of the same challenge using the theory of subjective logic. Sensoy et al. treated the predictions of an NN as subjective opinions and introduced a novel loss function that collects the evidence leading to these opinions by using a deterministic NN by applying a Dirichlet distribution to the class probabilities. The analysis carried out by Sensoy et al. emphasizes the peculiarities of their novel loss function, which enhances uncertainty estimation.

5. Further Studies of UQ Methods

In this section, we discuss few other UQ methods used in machine and deep learning algorithms to estimate uncertainty. In this regard, we present a summary of the proposed methods but not the theoretical parts. Due to the page limitation and large number of references, we are not able to review all the details of the methods. For this reason, readers can find more details of each method in the references if needed.

OoD is a common error that appears in machine and deep learning systems when the training data have different distributions. To address this issue, Ardywibowo et al. [195] introduced a new UA architecture called *Neural Architecture Distribution Search* (*NADS*). The proposed NADS finds an appropriate distribution of different architectures that perform significantly well on a specified task. A single block diagram for searching the space in the architecture is presented in Fig. 16.

Unlike previous architecture design methods, NADS enables the recognition of common blocks among all UA architectures. On the other hand, the cost functions for an uncertainty-oriented NN are not always convergent. Moreover, an optimized prediction interval (PI) is not always generated by converged NNs. The convergence of training is uncertain, and the NNs are not customizable for such cost functions. To construct the optimal PIs, Kabir et al. [196] presented a smooth customizable cost function to develop the optimal PIs to construct NNs. The PI coverage probability (PICP), PI failure distances and optimized average width of PIs were computed to lessen the variation in the quality of PIs, increase the convergence probability and speed up the training. They tested their method on electricity demand and wind power generation data. In the case of non-Bayesian deep neural classification, uncertainty estimation methods introduced biased estimates for instances whose predictions are highly accurate. They argued that this limitation occurred because of the dynamics of training with SGD-like optimizers and that it possessed similar characteristics, such as overfitting. Geifman et al. [197] proposed an uncertainty estimation method that computed the uncertainty of highly confident points by utilizing snapshots of the trained model before

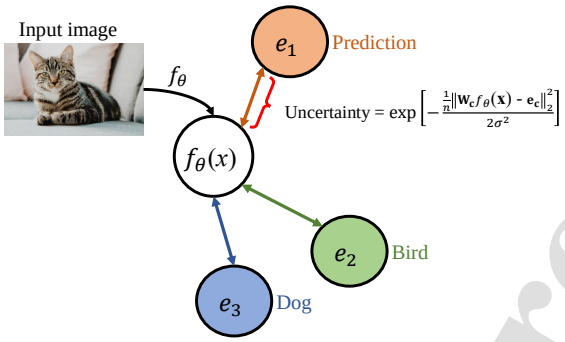


Fig. 17: A general view of the DUQ architecture which is reproduced based on [199, 200].

their approximations were jittered. The proposed algorithm outperformed all well-known techniques. In another work, Tagasovska et al. [198] proposed single-model estimates for DNNs of epistemic and aleatoric uncertainty. They suggested a loss function called simultaneous quantile regression (SQR) to obtain the conditional quantiles of a target variable to assess aleatoric uncertainty. Well-calibrated prediction intervals could be derived by using these quantiles. They devised orthonormal certificates (OCs), a collection of non-constant functions that mapped training samples to zero to estimate epistemic uncertainty. The OoD examples were mapped by these certificates to nonzero values.

Van Amersfoort et al. [199, 200] presented a method to find and reject distribution data points for training a deterministic deep model with a single forward pass at test time. They exploited the ideas of RBF networks to devise deterministic UQ (DUQ), which is presented in Fig. 17. They scaled the training in this method with a centroid updating scheme and a new loss function. Their method could detect OoD data consistently by utilizing a gradient penalty to track changes in the input. Their method was able to enhance DEs and scaled well to huge databases.

Tagasovska et al. [201] demonstrated frequentist estimates of epistemic and aleatoric uncertainty for DNNs. They proposed a loss function, simultaneous quantile regression, to estimate all the conditional quantiles of a given target variable in case of aleatoric uncertainty. Well-calibrated prediction intervals could be measured by using these quantiles. They proposed a collection of non-trivial diverse functions that mapped all training samples to zero and doubled as training certificates for the estimation of epistemic uncertainty. The certificates signaled high epistemic uncertainty by mapping OoD examples to nonzero values. By using Bayesian deep networks, it is possible to know what DNNs do not know in domains where safety is a major concern. Flawed decisions may lead to severe penalties in many domains such as autonomous driving, security and medical diagnosis. Traditional approaches are incapable of scaling to complex large neural networks. Mobiny et al. [202] proposed an approach imposing a Bernoulli distribution on the model weights to approximate Bayesian inference for DNNs. Their framework, called MC-DropConnect, represented model uncertainty through small alterations in the model structure or computed cost. They validated their technique on various datasets and architectures for semantic segmentation and classification tasks. They also introduced a novel uncertainty quantification metric. Their experimental results showed considerable enhancements in uncertainty estimation and prediction accuracy compared to prior approaches.

Uncertainty measures are crucial estimating tools in the machine learning domain that can evaluate the similarity and dependence between two feature subsets and can be utilized to verify the importance of features in clustering and classification algorithms. There are few uncertainty measurement tools that estimate a feature subset, including rough entropy, information entropy, roughness, and accuracy, in classical rough sets. These measures are not appropriate for real-valued datasets and are relevant to discrete-valued information systems. Chen et al. [203] proposed the neighborhood rough set model. In their approach, each object is related to a neighborhood subset, dubbed a neighborhood granule. Different uncertainty measures for neighborhood granules were introduced: information granularity, neighborhood entropy, information quantity, and neighborhood accuracy. Furthermore, they confirmed that these measures of uncertainty ensure monotonicity, invariance and non-negativity. In neighborhood systems, their experimental results

and theoretical analysis demonstrated that information granularity, neighborhood entropy and information quantity performed better than the neighborhood accuracy measure. On the other hand, reliable and accurate machine learning systems depend on techniques for reasoning under uncertainty. UQ is provided by a framework using Bayesian methods. However, Bayesian uncertainty estimations are often imprecise because of the use of approximate inference and model misspecification. Kuleshov et al. [204] devised a simple method for calibrating any regression algorithm; it was guaranteed to provide calibrated uncertainty estimates that have enough data when used on probabilistic and Bayesian models. They assessed their technique on recurrent, feedforward NNs and Bayesian linear regression, and they located the outputs of well-calibrated credible intervals while enhancing the performance on model-based RL and time series forecasting tasks.

Gradient-based optimization techniques have showed efficacy in learning overparameterized and complex NNs from non-convex objectives. Nevertheless, generalization to DNNs, the induced training dynamics, and the specific theoretical relationships between gradient-based optimization methods are still unclear. Rudner et al. [205] examined the training dynamics of overparameterized NNs under natural gradient descent. They demonstrated that the discrepancy between functions obtained from nonlinearized and linearized natural gradient descent is smaller than that of standard gradient descent. They showed empirically that there was no need to formulate a limit argument about the width of the neural network layers, as the discrepancy was small for overparameterized neural networks. Finally, they demonstrated that the discrepancy was small on a set of regression benchmark problems, and their theoretical results were steady, with an empirical discrepancy between the functions obtained from nonlinearized and linearized natural gradient descent. Patro et al. [206] devised gradient-based certainty estimates with visual attention maps. They resolved visual question answering tasks. The gradients for the estimates were enhanced by incorporating probabilistic deep learning techniques. There are two key advantages to this method: 1. the enhancement in obtaining the certainty estimates correlated better with misclassified samples, and 2. the SOTA results obtained by improving the attention maps correlated with human attention regions. The enhanced attention maps consistently improved different techniques for visual question answering. Improved certainty estimates and explanations of deep learning techniques could be achieved through the presented method. They provided empirical results on all benchmarks for the visual question answering task and compared them with those of standard techniques.

BNNs have been used as a solution for NN predictions, but it is still an open challenge to specify their priors. An independent normal prior in the weight space leads to weak constraints on the function posterior, which permit it to generalize in unanticipated ways on inputs outside of the training distribution. Hafner et al. [14] presented noise contrastive priors (NCPs) to estimate consistent uncertainty. The prime goal was to train the model for data points outside of the training distribution to output an elevated uncertainty. The NCPs relied on input priors that included noise in the inputs of the current mini-batch and an output prior that was an extensive distribution set for these inputs. The NCPs restricted overfitting outside of the training distribution and produced convenient uncertainty estimates for AL. BNNs with latent variables are flexible and scalable probabilistic models. They can record complex noise patterns in the data by using latent variables, and uncertainty is accounted for by network weights. Depeweg et al. [207] exhibited decomposition and divided uncertainty into aleatoric and epistemic types for decision support systems. This enabled them to detect informative points for AL of functions with bimodal and heteroscedastic noise. They further described a new risk-sensitive condition to recognize policies for RL that balanced noise aversion, model bias and the expected cost by applying decomposition.

Uncertainty modeling in DNNs is an open problem despite advancements in the area. BNNs, where the prior over the network weights is a design choice, are a powerful solution. Frequently, a normal or other distribution supports sparsity. The prior is agnostic to the generative process of the input data. This may lead to unwarranted generalization for out-of-distribution tested data. Rohekar et al. [208] suggested a confounder for the relation between the discriminative function and the input data given the target label. They proposed modeling the confounder by sharing neural connectivity patterns between discriminative and generative networks. Hence, a novel deep architecture was framed, where networks were coupled into a compact hierarchy and sampled from the posterior of local causal structures (see Fig. 18).

They showed that sampling networks from the hierarchy, an efficient technique, was proportional to their posteriors and that different types of uncertainties could be estimated. It is a challenging job to learn unbiased models from imbalanced datasets. The generalization of learned boundaries to novel test examples is hindered

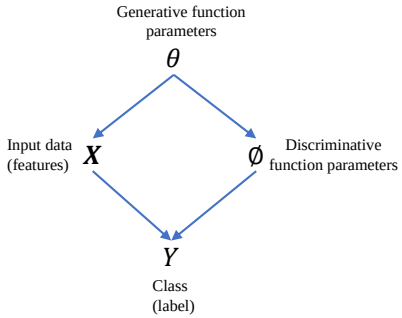


Fig. 18: A causal view demonstrating the main assumptions made by Rohekar et al. [208] (this figure is reproduced based on the reference).

by concentrated representation in the classification space for rare classes. Khan et al. [209] showed that the difficulty level of individual samples and the rarity of classes had direct correlations with Bayesian uncertainty estimates. They presented a new approach for uncertainty-based class imbalance learning that exploited two insights: 1. In rare (uncertain) classes, the classification boundaries should be broadened to avoid overfitting and improve its generalization; and 2. a sample's uncertainty is defined by a multivariate Gaussian distribution with a covariance matrix and a mean vector that models each sample. Individual samples and their distributions in the feature space should be addressed with the learned boundaries. Class and sample uncertainty information was used to obtain generalizable classification techniques and robust features. They formulated a loss function for max-margin learning based on a Bayesian uncertainty measure. Their technique exhibited key performance enhancements on six benchmark databases for skin lesion detection, digit and object classification, attribute prediction and face verification.

NNs do not measure uncertainty meaningfully, as they tend to be overconfident on incorrectly labeled, noisy or unseen data. Variational approximations such as multiplicative normalizing flows and BBB are utilized by BDL to overcome this limitation. However, current methods have shortcomings regarding scalability and flexibility. Pawlowski et al. [210] proposed the novel technique of variational approximation, termed Bayes by hypernet (BbH), which deduced hypernetworks as implicit distributions. It naturally scaled to deep learning architectures and utilized neural networks to model arbitrarily complex distributions. Their method was robust against adversarial attacks and yielded competitive accuracies. On the other hand, although a significant increase in prediction accuracy resulted for deep learning models, it was accompanied by an increase in the cost of rendering predictions. Wang et al. [211] speculated that for many real-world inputs, deep learning models created recently tend to "overthink" on simple inputs. They proposed the I Don't Know (IDK) prediction cascade approach to create a set of pretrained models systematically without a loss in prediction accuracy to speed up inference. They introduced two search-based techniques for producing a new cost-aware objective as well as cascades. Their IDK cascade approach can be adopted in a model without further model retraining. They tested its efficacy on a variety of benchmarks.

Yang et al. [212] proposed a deep learning approach for propagating and quantifying uncertainty in models inspired by the nonlinear differential equations utilized by physics-informed neural networks. Probabilistic representations for the system states were produced by latent variable models, while physical laws described by partial differential equations were satisfied by constraining their predictions. An adversarial inference method for training them on data was also proposed. A regularization approach for efficiently training deep generative models was provided by such physics-informed constraints. Surrogates of physical models were given, in which there was usually little training of datasets and the cost of data acquisition was high. The outputs of physical systems were characterized by the framework due to noise in their observations or randomness in their inputs, which bypassed the need for sampling costly experiments or numerical simulators. They proved the efficacy of their method via a series of examples that demonstrated uncertainty propagation in nonlinear conservation laws and the detection of constitutive laws. For autonomous driving, 3D scene flow estimation techniques generate 3D motion for a scene and 3D geometry. Brickwedde et al. [213] devised a new monocular 3D scene flow estimation technique, dubbed Mono-SF, that assessed both the motion

of a scene and the 3D structure by integrating single-view depth information and multiview geometry. A CNN algorithm termed ProbDepthNet was devised for combining single-view depth in a statistical manner. The new recalibration technique of ProbDepthNet was presented for regression problems to guarantee well-calibrated distributions. The ProbDepthNet design and Mono-SF method proved their efficacy in comparison to the SOTA approaches.

Mixup is a DNN training technique in which extra samples are produced during training by convexly integrating random pairs of images and their labels. The method has demonstrated its effectiveness in improving image classification performance. Thulasidasan et al. [214] investigated the predictive uncertainty and calibration of models trained with mixup. They revealed that DNNs trained with mixup were notably better calibrated than those trained by typical techniques. They tested their technique in large datasets and observed that it was less likely to yield overconfident predictions using random noise and OoD data. Label smoothing in mixup-trained DNNs played a crucial role in enhancing calibration. They concluded that training with hard labels caused overconfidence in neural networks. The transparency, fairness and reliability of the methods can be improved by explaining black-box machine learning models. Model robustness and user trust raised concerns, as the explanations of these models exhibited considerable uncertainty. Zhang et al. [215] illustrated three sources of uncertainty, viz., variation in the explained model credibility, variation with sampling proximity and randomness in the sampling procedure across different data points, by concentrating on a specific local explanation technique called local interpretable model-agnostic explanations (LIME). Even black-box models with high accuracy yielded uncertainty. They tested uncertainty with the LIME technique on two publicly available datasets and synthetic data.

For even small adversarial perturbations, the employment of DNNs in safety-critical environments is rigorously restricted. Sheikholeslami et al. [216] devised a randomized approach to identify such perturbations that dealt with minimum uncertainty metrics by sampling from the hidden layers during the DNN inference period. Adversarial corrupted inputs were identified by the sampling probabilities. Any pretrained DNN, with no additional training, could be exploited by a new detector of adversaries. The output uncertainty of the DNN, from the perspectives of the BNNs, could be quantified by choosing the units to sample per hidden layer, where layerwise components denoted the overall uncertainty. Low-complexity approximate solvers were obtained by simplifying the objective function. These approximations associated SOTA randomized adversarial detectors with the new approach in addition to delivering meaningful insights. Moreover, the consistency loss between various predictions under random perturbations is the basis of an effective strategy in semisupervised learning. In a successful student model, the teachers' pseudo-labels must have good quality; otherwise, the learning process will suffer. However, the prevailing models do not evaluate the quality of the teachers' pseudo-labels. Li et al. [217] presented a new certainty-driven consistency loss (CCL) that employed the predictive uncertainty information in the consistency loss to teach students from reliable targets dynamically. They devised two strategies, i.e., temperature CCL and filtering CCL, to either pay less attention to the uncertain data or filter out uncertain predictions in the consistency regularization. They termed it FT-CCL by integrating the two strategies to enhance the consistency learning approach. FT-CCL demonstrated robustness to noisy labels and enhancement on a semisupervised learning task. They presented a new mutual learning technique in which one student was detached from its teacher and gained additional knowledge with another student's teacher.

Englesson et al. [218] introduced a modified knowledge distillation method to achieve computationally competent uncertainty estimates with deep networks. They tried to obtain competitive uncertainty estimates for both out-of- and in-distribution samples. Their major contributions were as follows: 1. adapting and demonstrating the regularization effect of distillation, 2. presenting a new target teacher distribution, 3. enhancing OoD uncertainty estimates by a simple augmentation method, and 4. executing a broad set of experiments to shed light on the distillation method. On the other hand, well-calibrated uncertainty and accurate fully predictive distributions are provided by Bayesian inference. The high dimensionality of the parameter space limits the scaling of Bayesian inference methods to DNNs. Izmailov et al. [28] designed low-dimensional subspaces of parameter space that comprised diverse sets of high-performing approaches. They applied VI and elliptical slice sampling in the subspaces. Their method yielded well-calibrated predictive uncertainty and accurate predictions for both image classification and regression by exploiting Bayesian model averaging over the induced posterior in the subspaces.

Csáji et al. [219] introduced a data-driven strategy for uncertainty quantification of models based on kernel techniques. This method needed a few mild regularities in the computation of noise instead of distributional

assumptions such as dealing with exponential families or GPs. The uncertainty of the model could be estimated by perturbing the residuals in the gradient of the objective function. They devised an algorithm to make the model distribution-free and non-asymptotically guaranteed and to provide exact confidence regions for the noise-free and ideal depiction of the functions that they estimated. Regarding symmetric noise and the usual convex quadratic problems, the regions were star convex centered on a specified small estimate, and ellipsoidal outer approximations were also efficiently executed. On the other hand, the uncertainty estimates could be measured during the pretraining process. Hendrycks et al. [220] demonstrated that pretraining enhanced the uncertainty estimates and model robustness, although it might not improve the classification metrics. They showed the key gains from pretraining by performing empirical experiments on confidence calibration, OoD detection, class imbalance, label corruption and adversarial examples. Their adversarial pretraining method demonstrated approximately 10% enhancement over existing methods in adversarial robustness. Pretraining without task-specific techniques highlighted the need for pretraining and surpassed the SOTA performance when examining future techniques for uncertainty and robustness.

Trustworthy confidence estimates are required from predictive models in high-risk domains. Rigid variational distributions utilized for tractable inference that err on the side of overconfidence suffer from deep latent variable models. Veeling et al. [221] devised stochastic quantized activation distributions (SQUAD), which executes a tractable yet flexible distribution over discretized latent variables. The presented technique is sample efficient, self-normalizing and scalable. Their method yielded predictive uncertainty of high quality, learned interesting nonlinearities, and fully used the flexible distribution. Multitask learning (MTL) is another domain in which the impact of the importance of uncertainty methods can be considered. For example, MTL demonstrated its efficacy for MR-only radiotherapy planning, as it can jointly automate the contour of organs at risk - a segmentation task – and simulate a synthetic CT (synCT) scan - a regression task - from MRI scans. Bragman et al. [222] suggested utilizing a probabilistic deep learning technique to estimate parameter and intrinsic uncertainty. Parameter uncertainty was estimated through approximate Bayesian inference, while intrinsic uncertainty was modeled using a heteroscedastic noise technique. This led to an approach for uncertainty measurement overprediction of the tasks and the data-driven adaptation of task losses on a voxelwise basis. This method demonstrated competitive performance in the segmentation and regression of prostate cancer scans.

As a matter of fact, Gaussian Processes (GP) [223] is a powerful technique used for quantifying uncertainty. However, it is a complex task to form a Gaussian approximation to a posterior distribution even in the context of uncertainty estimation in huge deep learning models. In such scenarios, the prevailing techniques generally lead to a diagonal approximation of the covariance matrix in spite of executing low-uncertainty estimates with these matrices. Mishkin et al. [224] designed a novel stochastic, low-rank, approximate natural gradient (SLANG) technique for VI in huge deep models to address this issue. Their technique computed a "diagonal plus low-rank" structure based on the backpropagated gradients of the network log-likelihood. Their findings indicate that the proposed technique shapes the Gaussian approximation to the posterior distribution. In fact, the safety of AI systems can be enhanced by estimating the uncertainty in predictions. Such uncertainties arise due to a distributional mismatch between the training and test data distributions, irreducible data uncertainty and uncertainty in the model parameters. Malinin et al. [225] devised a novel framework for predictive uncertainty dubbed prior networks (PNs), which model distributional uncertainty explicitly. This was achieved by parameterizing a prior distribution over predictive distributions. Their work aimed at determining uncertainty in classification and scrutinized PNs on the tasks of recognizing OoD samples and identifying misclassifications on the CIFAR-10 and MNIST datasets. The empirical results indicated that PNs, unlike non-Bayesian methods, could successfully discriminate between distributional and data uncertainty.

It has been observed in recent research findings that CNN models for image classification show overlapping adversarial vulnerabilities. Adversarial training eradicates the vulnerability in a single model by forcing it to learn robust features. This process is rigid and suffers from substantial loss on clean data accuracy. Alternatively, ensemble techniques have been presented to induce submodels with diverse outputs against a transfer adversarial example. Only a minor clean accuracy drop is perceived in this process. However, earlier ensemble training techniques are not efficient in inducing such diversity and are thus unproductive in reaching robust ensembles. Yang et al. [226] devised DVERGE, which segregated the adversarial vulnerability of each submodel by distilling non-robust features and diversified the adversarial vulnerability to induce diverse outputs against a transfer attack. The new diversity metric and training procedure permitted

DVERGE to yield higher robustness against transfer attacks. Moreover, epistemic uncertainty is the part of out-of-sample prediction error that is due to the absence of knowledge of the learner. Earlier works focused on model variance, whereas Jain et al. [227] introduced a principled approach for estimating epistemic uncertainty by learning to predict the generalization error and subtracting an estimate of aleatoric uncertainty, i.e., intrinsic unpredictability. The estimator of epistemic uncertainty included the effect of model bias and could be utilized in nonstationary learning environments arising in reinforcement learning or active learning. In addition to showcasing these properties of direct epistemic uncertainty prediction (DEUP), they showed its benefits against prevailing techniques for uncertainty estimation on downstream tasks, including reinforcement learning and sequential model optimization.

A few more studies have studied the importance of UQ methods in different research domains. For example, Daxberger et al. [228] presented inference performed over only a small subset of the model parameters while keeping all others as point estimates. This enabled the utilization of expressive posterior approximations that were intractable in the full model. They devised, in particular, a scalable and practical Bayesian deep learning technique that trained a point estimate and then inferred a full covariance Gaussian posterior approximation over a subnetwork. It should be noted that restrictive approximations are needed for scaling Bayesian inference to the large parameter spaces of DNNs. Additionally, graph NNs (GNNs) do not usually consider the various types of uncertainties associated with class probabilities to mitigate risk that increases misclassification under uncertainty in real life. Zhao et al. [229] presented a subjective Bayesian deep learning approach dealing with different types of uncertainties for classification predictions by leveraging the learning and powerful modeling capabilities of GNNs. They considered the predictions of a subjective Bayesian GNN (S-BGNN) to be the nodes' multinomial subjective opinions in a graph based on Dirichlet distributions, where each belief mass was a belief probability of a class. By gathering evidence from the given labels of training nodes, the S-BGNN model was devised for predicting the probabilities of each class and discovering out-of-distribution data. Finally, neural processes (NPs) established variational approximate models for stochastic processes with promising properties in UQ and computational efficiency. These processes utilized neural networks with latent variable inputs to induce predictive distributions. Although target-specific local variation may sometimes be vital, the expressiveness of vanilla NPs is inadequate, as they apply only a global latent variable. To address this issue, Wang et al. [230] exploited NPs systematically and proposed a novel variant of NP modeling that they dubbed the doubly stochastic variational neural process (DSVNP). Their model integrated the local latent variables and global latent variable for prediction.

Investigation of uncertainty in non-Bayesian NN models is another important research topic. For this reason, Amini et al. [231] presented a technique for training non-Bayesian neural networks to compute a continuous target as well as its associated evidence to learn both the epistemic and aleatoric uncertainty. They achieved this by placing evidential priors over the original Gaussian likelihood function and training the neural networks to infer the hyperparameters of the evidential distribution. Their method accomplished scalable and efficient uncertainty learning, as their method did not depend on OoD examples or sampling during inference for training. Moreover, the error prediction methods using auxiliary regression models have also been used for error prediction in the literature. Predicting the errors that may occur and the reliably evaluating model confidence in deep learning are vital elements of demonstrating safety for model deployment. Tsiligkaridis [232] first showcased that uncertainty-aware deep Dirichlet neural networks delivered enhanced separation between the confidence of incorrect and correct predictions in the true class probability (TCP) metric. Second, as the true class was not known at the test time, a novel criterion was introduced for learning the true class probability by matching the prediction confidence scores while taking TCP and imbalance constraints into account for obtaining correct predictions and failures. In another study, Corbiere et al. [233] expressed that the confidence of DDNs and the prediction of their failures are of key importance for the practical application of these methods. In this regard, the authors showed that the TCP (*True Class Probability*) is more suitable than the MCP (*Maximum Class Probability*) for failure prediction in deep learning. For instance, in their model: $\mathbb{R}^d \times \mathcal{Y} \rightarrow \mathbb{R}$, and:

$$(x, y^*) \rightarrow P(Y = y^*|w, x), \tag{36}$$

where $x_i \in \mathbb{R}^d$ represents a d-dimensional feature and $y_i^* \in \mathcal{Y} = \{1, ..., K\}$ is its correct class. Then, they introduced a new normalized type of TCP confidence criterion, which is as follows:

$$TCP^r(x, y^*) = \frac{P(Y = y^*|w, x)}{P(Y = \hat{y}|w, x)}. \tag{37}$$

A general view of the model proposed in [233] is illustrated in Fig. 19.

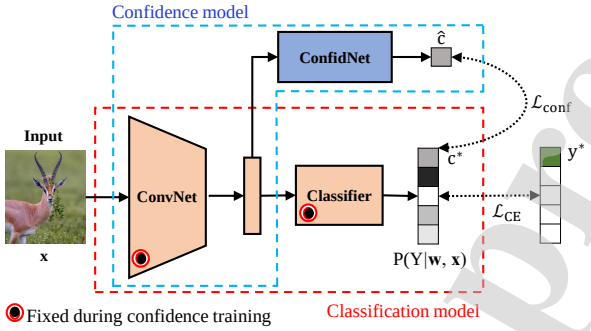


Fig. 19: A schematic view of the TCP model, which is reproduced based on [233].

More information on further UQ methods can be found in Tables B.3 and B.4 in the Appendix.

5.1. Deep Gaussian Processes

Deep Gaussian processes (DGPs) [234, 235, 236, 237, 238, 239, 240] are effective multilayer decision making models that can accurately model uncertainty. They represent a multilayer hierarchy of Gaussian processes (GPs) [241, 242]. A GP is a nonparametric type of Bayesian model that encodes the similarity between samples using a kernel function. It represents distributions over the latent variables with respect to the input samples as a Gaussian distribution $f_x \sim \mathcal{GP}(m(x), k(x, x'))$. Then, the output y is distributed based on a likelihood function $y|f_x \sim h(f_x)$. However, conventional GPs cannot effectively scale to large datasets. To address this issue, inducing samples can be used. As such, the following variational lower bound can be optimized:

$$\log p(Y) \geq \sum_{y,x \in Y,X} \mathbb{E}_{q(f_x)}[\log p(y|f_x)] - \mathrm{KL}(q(fz)\|p(fz)), \tag{38}$$

where Z and $q(f_x)$ are the location of the inducing samples and the approximated variation with respect to the distribution of f_x, respectively.

Oh et al. [243] proposed hedged instance embedding (HIB), which hedges the position of each sample in the embedding space, to model the uncertainty when the input sample is ambiguous. As such, the probability of two samples matching was extended to stochastic embedding, and MC sampling was used to approximate it. Specifically, a mixture of C Gaussians was used to represent the uncertainty. Havasi et al. [244] applied SGHMC into DGPs to approximate the posterior distribution. They introduced moving window MC expectation maximization to obtain the maximum likelihood to deal with the problem of optimizing a large number of parameters in DGPs. Maddox et al. [245] used stochastic weight averaging (SWA) [246] to build a Gaussian-based model to approximate the true posterior. Later, they proposed SWA-G [247], SWA-Gaussian, to model Bayesian averaging and estimate uncertainty. Most weight perturbation-based algorithms suffer from the high variance of gradient estimation due to sharing the same perturbations among all samples in a mini-batch. To alleviate this problem, flipout [248] was proposed. Flipout samples the pseudo-independent weight perturbations for each input to decorrelate the gradient within a mini-batch. It is able to reduce the variance and computational time in training NNs with multiplicative Gaussian perturbations. Despite the success of DNNs in dealing with complex and high-dimensional image data, they are not robust to adversarial examples [249]. Bradshaw et al. [250] proposed a hybrid model of GP and DNNs (GPDNNs) to deal with the uncertainty caused by adversarial examples (see Fig. 20).

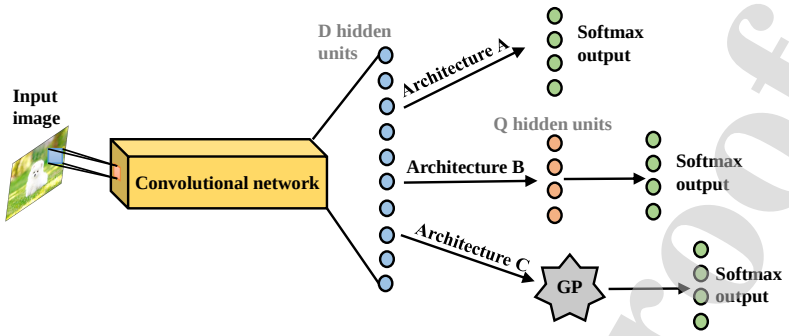


Fig. 20: A general Gaussian-based DNN model proposed by Bradshaw et al. [250], which is reproduced based on the given reference.

Choi et al. [251] proposed a Gaussian-based model to predict the localization uncertainty in YOLOv3 [252]. To this end, they applied a single Gaussian model to the bbox coordinates of the detection layer. Specifically, the coordinates of each bbox were modeled as the mean (μ) and variance ($\sum$) to predict the uncertainty of the bbox. Khan et al. [253] proposed a natural gradient-based algorithm for Gaussian mean-field VI. A Gaussian distribution with diagonal covariances was used to estimate the probability. The proposed algorithm was implemented by the Adam optimizer. To achieve this, the network weights were perturbed during gradient evaluation. In addition, they used a vector to adapt the learning rate to estimate uncertainty.

Sun et al. [254] considered the structural information of the model weights. They used the matrix variate Gaussian (MVG) [255] distribution to model structured correlations in the weights of DNNs and introduced a reparametrization for the MVG posterior to make posterior inference feasible. The resulting MVG model was applied to a probabilistic BP framework to estimate the posterior inference. Louizos and Welling [256] used an MVG distribution to estimate the weight posterior uncertainty. They treated the weight matrix as a whole rather than treating each component of the weight matrix independently. As mentioned earlier, GPs have been widely used for UQ in deep learning methods. Van der Wilk et al. [257], Blomqvist et al. [258], Tran et al. [259], Dutordoir et al. [260] and Shi et al. [261] introduced convolutional structure into GP.

In another work, Atanov et al. [262] introduced a probabilistic model and showed that the batch normalization (BN) approach can maximize the lower bound of its related marginalized log-likelihood. Since inference was computationally inefficient, they proposed stochastic BN (SBN) for approximating the appropriate inference procedure as an uncertainty estimation method. Moreover, induced noise is generally employed to capture uncertainty, determine overfitting and slightly improve performance via test-time averaging, whereas ordinary stochastic neural networks typically depend on the expected values of their weights to formulate predictions. Neklyudov et al. [263] proposed a different kind of stochastic layer called a variance layer. It is parameterized by its variance, and each weight of a variance layer obeys a zero-mean distribution. This implies that each object is denoted by a zero-mean distribution in the space of the activations. They demonstrated that these layers present a strong defense against adversarial attacks and could serve as a crucial exploration tool in reinforcement learning tasks.

5.2. Uncertainty Quantification in the Traditional Machine Learning Domain Using Ensemble Techniques

It is worth noting that UQ in traditional machine learning algorithms has been studied extensively using different ensemble techniques and a few more UQ methods (e.g., please see [264]) or other UQ methods in classification problems [265] in the literature. However, due to page limitations, we only summarize some of the ensemble techniques (as UQ methods) used in the traditional machine learning domain. For example, Tzelepis et al. [264] proposed a maximum margin classifier to deal with uncertainty in input data. The proposed model was applied to classification tasks using the support vector machine (SVM) algorithm with multidimensional Gaussian distributions. The proposed model, named SVM with Gaussian sample uncertainty (SVM-GSU), is illustrated in Fig. 21:

In another work, Pereira et al. [266] examined various techniques for transforming classifiers into uncer-

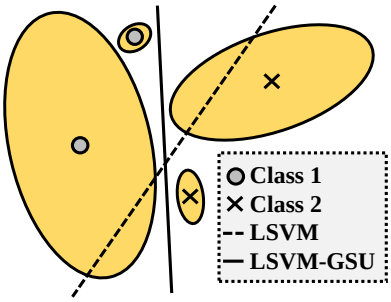


Fig. 21: A schematic view of SVM-GSU which is reproduced based on [264].

tainty methods whereby predictions are harmonized with probability estimates through their uncertainty. They applied various uncertainty methods: Venn-ABERS predictors, conformal predictors, Platt scaling and isotonic regression. Partalas et al. [267] presented a novel measure called uncertainty-weighted accuracy (UWA) for ensemble pruning through directed hill climbing that addressed the uncertainty of the current ensemble decision. The experimental results demonstrated that the new measure for pruning a heterogeneous ensemble had significantly enhanced accuracy compared to that of baseline methods and other SOTA measures. Peterson et al. [268] exploited different types of errors that might creep into atomistic machine learning and addressed how uncertainty analysis validated machine learning predictions. They applied a bootstrap ensemble of neural network-based calculators and showed that the width of the ensemble can represent an approximation of the uncertainty.

6. Applications

In this section, we discuss a few of the most important applications of different UQ techniques used in machine and deep learning methods. In this regard, we first summarize the application of UQ techniques in image processing and computer vision followed by medical image analysis. Then, we show how UQ techniques have been applied to NLP, and some additional (further) applications.

6.1. Image Processing and Computer Vision

Currently, deep learning algorithms are being widely used to map high-dimensional data to output arrays; however, these mappings can be inaccurate in many cases, such as when two African Americans were incorrectly identified as gorillas in an image classification system [269], which led to racial discrimination [270]. Therefore, it is important to take uncertainty into account when predictions are made by deep learning-based computer vision algorithms. A number of studies have addressed uncertainty in deep learning algorithms for various applications during the years 2016 to 2020, including but not limited to image/video retrieval [271, 272], depth estimation [273, 274], object detection [275, 276, 277], semantic segmentation and scene understanding [278, 279, 280, 281, 10], optical flow estimation and motion prediction [282, 283, 284], human pose estimation and pedestrian localization [285, 286, 287], person reidentification and face recognition [288, 289, 290], camera relocalization [291], and avoiding adversarial attacks [292, 293].

In fact, most research studies in deep learning applications concentrate on prediction accuracy. Unlike these studies, untangling the complexity of various DNNs and addressing uncertainty for a variety of computer vision tasks has attracted significant interest [294]. There has still been a good record in using BNNs and MC dropout for uncertainty estimation in deep learning architectures. Nine studies have reported MC dropout as the most effective UQ technique [10, 44, 270, 282, 284, 286, 291, 293, 295], which is applicable to various deep learning architectures. Kendall et al showed that the uncertainty of their Bayesian convolutional neural network model came from the appearance and pose dissimilarity of images to the training set and could be used to estimate the model's relocalization uncertainty, which improved localization accuracy on a large outdoor dataset [291]. The same authors developed a measure of model uncertainty by MC sampling with dropout and enhanced the semantic segmentation performance compared to that of the SOTA methods in 2016 [44]. Eldesokey et al. [296] proposed a UA model for CNNs and tested it on the KITTI dataset. The

proposed model identified the disturbed measurements of the input data after learning an input confidence estimator by a self-supervised procedure using normalized CNNs (NCNNs).

Indeed, epistemic uncertainty estimation is a challenging problem, and while several scalable techniques have appeared lately, no widespread assessment has been carried out in a real-world setting. Gustafsson et al. [297] devised a comprehensive assessment framework for scalable epistemic uncertainty estimation techniques in deep learning. Their framework tested for the robustness needed in real-world computer vision applications. They also utilized their framework to conclusively and extensively compare two scalable techniques: MC dropout and ensembling. Postels et al. [281] proposed a sampling-free method for estimating the epistemic uncertainty of a neural network. Epistemic uncertainty is crucial in safety-critical applications, since it denotes the reliability of predictions using new data. Their prime contribution was the approximation of the epistemic uncertainty estimated by these techniques, which did not necessitate sampling, thus remarkably reducing the computational overhead. They used their method in volumetric visual tasks to showcase the advantages of their techniques in terms of computational overhead as well as uncertainty estimates.

Cai et al. [298] worked on the hand segmentation generalization issue without using segmentation labels in the target domain. They designed a Bayesian CNN-based model adaptation approach for hand segmentation, which devised and considered two vital factors: 1) general information about hand shapes shared across domains and 2) prediction uncertainty when the model is used in a new domain. Accordingly, they introduced an iterative self-training strategy for hand segmentation in a novel domain, which was directed by the model uncertainty approximated by a Bayesian CNN. However, Bayesian techniques have not been exploited extensively for 3D modalities such as the point clouds often utilized for autonomous systems and robots. Bhandary et al. [299] examined three uncertainty quantification techniques, viz., MC-DropConnect, MC dropout and DE, on the DarkNet21Seg 3D semantic segmentation model and analyzed the impact of different parameters such as drop probability values on task performance, the number of models in ensembles and forward passes and the uncertainty estimate quality. They demonstrated that DEs generated better results than other methods in terms of uncertainty metrics and performance.

Weakly supervised semantic segmentation using image-level labels is accomplished by acquiring object response maps. However, the prevailing techniques only depend on the classifier that can result in a response map attending to discriminative object regions, as the network does not require seeing the complete object in order to optimize the classification loss. Chang et al. [300] introduced a principled and end-to-end trainable approach to let the network pay attention to other parts of the object while generating a more uniform and complete response map. They specifically proposed a mixup data augmentation strategy for the classification network and devised two uncertainty regularization terms to better cooperate with the mixup scheme. More information regarding different UQ methods applied in computer vision and image processing tasks is given in Table C.5 in the Appendix.

6.1.1. Uncertainty in Semantic Segmentation

Semantic segmentation is an important computer vision and image processing task that assigns semantic labels from various predefined classes by considering all pixels of any input image [301]. Indeed, semantic segmentation techniques can generally be divided into two main categories: (1) those that obtain accurate results with slow inference, and (2) those that perform real-time inference by sacrificing performance for speed [301]. Measuring and quantifying uncertainty in semantic segmentation plays a significant role in having trustworthy outcomes. Regarding this point, a broad range of UQ methods have been used widely in semantic segmentation in the literature [278, 279, 301, 302]. To cover this important research domain, we briefly summarize the more recent studies on semantic segmentation with the application of UQ methods.

Deep learning has gained much attention for semantic segmentation in particular, with BDL utilized to obtain uncertainty maps from deep models when predicting semantic classes. A BDL method requires new metrics to examine whether it produces better uncertainty estimates than other methods. Mukhoti and Gal [10] introduced three novel metrics to evaluate BDL models devised specifically for the task of semantic segmentation. They modified DeepLab-v3+ and created its Bayesian counterpart using concrete dropout and MC dropout as inference techniques. They compared and tested these two inference methods on the Cityscapes dataset using their proposed metrics. In another study, DeVries and Taylor [55] estimated uncertainty to showcase two significant outputs in deep learning-based segmentation. The first one generated spatial uncertainty maps, from which a clinician could observe why and where the system failed. The other quantified an image-level prediction of failure, which was helpful for separating precise cases and removing

them from automated pipelines.

Zheng and Yang [303] focused on the unsupervised domain adaptation of transferring knowledge from the source domain to the target domain in the context of semantic segmentation. The pseudo-labels of the target-domain data are usually predicted by a model trained on the source domain. The generated labels comprise incorrect predictions due to the discrepancy between the test domain and training domain. To address this issue, the authors estimated the prediction uncertainty during training to fix pseudo-label learning for unsupervised semantic segmentation adaptation. The uncertainty of the prediction and the semantic segmentation prediction were estimated from the input image. On the other hand, many adversarial-based unsupervised domain adaptation (UDA) methods suffer from high-instability training and have to carefully tune the optimization procedure. Some nonadversarial UDA techniques exploit consistency regularization on the target predictions of a student model and a teacher model under various perturbations, where the teacher shares the same architecture as the student and is updated by the exponential moving average of the student. However, these models undergo negative transfer resulting from either an unreasonable teacher model or an error-prone discriminator network. Zhou et al. [304] presented an uncertainty-aware consistency regularization technique for cross-domain semantic segmentation. More reliable and meaningful knowledge from the teacher model could be transferred to the student model by employing the latent uncertainty information of the target samples.

The training of different deep learning-based image segmentation models needs a considerable amount of manual annotations. To mitigate this problem, Ravanbakhsh et al. [305] introduced a technique based on a conditional generative adversarial network (cGAN) that approached segmentation in a human-in-the-loop manner with semisupervised setup. More specifically, they applied a discriminator to detect unreliable slices for which expert annotation was needed and utilized the generator in the GAN to synthesize segmentations on unlabeled data for which the model was confident. The reliability of predictions is of greatest interest in the semantic segmentation of street scenes with neural networks. On the other hand, the evaluation of NNs by means of uncertainty is a typical measure to prevent safety problems. As video streams are available in applications such as automated driving, Maag et al. [306] proposed a time-dynamic framework to examine uncertainties and evaluate the prediction quality of neural networks. They tracked segments over time and collected aggregated metrics for each segment, thus obtaining a series of metrics with which they assessed the prediction quality. In Table 3, we summarize a few more UQ methods used in semantic segmentation.

Table 3: Further information on some UQ methods used in semantic segmentation.

Study	Year	Data source	Method	UQ method	Code
Postels et al. [281]	2019	Synthetic data and UCI regression datasets	Bayesian SegNet	OUR (our approximation)	✓
Ge et al. [307]	2018	Pascal VOC 2007 and Pascal VOC 2012	CNN	Pixel Labeling with Uncertainty	×
Blum et al. [308]	2019	FS Static, FS Web, and FS Lost & Found	Learned Embedding Density	NLL (minimizing the negative loglikelihood)	×
Rottmann and Marius et al. [309]	2019	Cityscapes	DeepLabv3+ MobilenetV2	Softmax outputs	✓
Hochgeschwender et al. [299]	2020	DarkNet21Seg	BNN	DEs	×
Chang et al. [300]	2020	PASCAL VOC 2012	Mixup-CAM	Uncertainty regularization	×
Hu et al. [310]	2020	ADE20K and Pascal-Context	ResNet-50	Bayesian uncertainty estimation	×
Xie et al. [311]	2020	CamVid and Cityscapes	DeepLabv3+ MobilenetV2	DEAL (Difficulty-awarE Active Learning)	×
Siddiqui et al. [312]	2020	SceneNet-RGBD, ScanNet, and Matterport3D	ViewAL	Viewpoint Entropy	✓

6.2. Medical Applications

Automated analysis of medical images came into existence as soon as it was possible to load and scan medical images into a computer [313]. At the outset, from the 1970s to the 1990s, medical image analysis was performed with the sequential application of low-level pixel processing (region growing, edge and line detector filters) and mathematical modeling (fitting lines, ellipses and circles) to build compound rule-based

systems that carried out particular tasks. This is analogous to the expert systems with many if-then-else statements that were popular in artificial intelligence in the same period. At the end of the 1990s, supervised methods, where training samples are used to develop a system, gradually became more popular in medical image analysis. Examples include active shape models, atlases, the concept of feature extraction and the use of statistical classifiers. This machine learning approach is still very popular and forms the foundation for various booming commercially available medical image analysis systems. Hence, there is a shift from systems that are devised entirely by humans to systems that are trained by computers utilizing example data from which feature vectors are derived. Computer algorithms establish the optimal decision boundary in a high-dimensional feature space. Both the monetary and ethical costs of erroneous predictions can be noteworthy in medicine, and the complexity of this issue leads to progressively more complex models. Although DL methods have achieved outstanding performance in medical image analysis, most of them have not been employed in highly automated disease monitoring systems due to the lack of reliability of the model [314]. For example, Dusenberry et al. [315] studied the role of model uncertainty strategies in the medical domain. They demonstrated that population-level metrics, such as calibration error, log-likelihood, AUC-ROC and AUC-PR, did not capture model uncertainty, and this was shown by applying RNN ensembles and different BRNNs. They showed that RNNs with only Bayesian embeddings could be a better way to address model uncertainty than ensembles. They further demonstrated that RNNs with only Bayesian embeddings yielded better results in model uncertainty than ensembles.

Well-annotated medical data is extremely expensive for conducting medical image segmentation. However, unlabeled data are a very appropriate solution that can be used in both semisupervised and unsupervised learning domains. As discussed earlier, Xia et al. [316] introduced the UMCT model as a semisupervised framework and tested it on various medical image datasets. They extended the Dice loss for uncertainty-weighted label fusion (ULF) as follows:

$$\mathcal{L}_{Dice} = \frac{1}{D}\Sigma_{d=0}^{D}\frac{2\Sigma_{i=1}^{N} y_i^d \hat{y}_i^d}{\Sigma_{i=1}^{N}(y_i^d)^2 + \Sigma_{i=1}^{N}(\hat{y}_i^d)^2}. \tag{39}$$

According to the obtained results, the proposed UMCT method outperformed the other applied methods on the same datasets. As a result, it was concluded that having a proper uncertainty method can assist in obtaining better medical image analysis performance.

Blood oxygen saturation (sO_2) measurement by optical imaging oximetry offers insight into local tissue metabolism and functions. Traditional methods for quantifying sO_2 suffer from uncertainties due to variations in the experimental conditions, systemic spectral bias, light spectral bias, tissue geometry and biological variability. Liu et al. [317] devised deep spectral learning (DSL), a novel data-driven approach to obtaining oximetry that was robust to experimental variation and facilitated uncertainty quantification for each sO_2 prediction. The predictions calculated by DSL were highly adaptive to depth-dependent backscattering spectra as well as to experimental variability. The DSL-predicted sO_2 demonstrated notably lower mean-square errors than the traditional least-squares fitting method. Inherent ambiguities cause many real-world vision problems. It is difficult to access, for example, which region contains cancerous tissue from a CT scan in clinical applications. Kohl et al. [318] devised a generative segmentation model based on the combination of a U-Net with a conditional VAE that is capable of generating a large number of plausible hypotheses. They exhibited it on a Cityscapes segmentation task and a lung abnormality segmentation task approach to regenerate all the possible segmentation variants as well as the frequency with which they outperformed the existing methods.

In another work, Araújo et al. [319] proposed an uncertainty-aware deep learning model (which was named DR—GRADUATE) for grading diabetic retinopathy (DR) using eye fundus images. To this end, they introduced a new Gaussian sampling technique on a multiple-instance learning (MIL) framework and used the proposed system as a second-opinion DR diagnostic system. UQ methods have also been used in the prostate cancer domain. Karimi et al. [320] studied prostate cancer using ultrasound images. In this regard, they proposed a robust and accurate deep learning (CNN) segmentation model. Moreover, due to the importance of uncertainty in medical image analysis, they computed the uncertainty as follows:

$$Q = 1 - \bar{p}^2 - (1 - \bar{p})^2, \tag{40}$$

where $\bar{p}$ is the average of the applied probability maps. The results confirmed that adding uncertainty resulted in having better prostate cancer segmentation outcomes. As discussed above, the MC dropout demonstrated impressive performance in quantifying uncertainty in deep learning methods. Combalia et al. [321] applied MC dropout in DNNs for UQ in dermoscopic (skin lesion) image classification. Their results indicated that using different uncertainty metrics was an appropriate solution in exploring difficult and OoD samples.

Cardiovascular disease detection by machine and deep learning is another research topic for the application of UQ methods. 2D echocardiography is a widely used imaging modality for cardiovascular diseases. Deep learning techniques have been widely used in 2D echocardiography for structural and functional assessment and automated view classification. Most of the models do not estimate uncertainty in this regard, which is crucial. Dahal et al. [322] compared three ensemble-based uncertainty techniques utilizing four different metrics to achieve insight into uncertainty modeling for left ventricular segmentation from ultrasound (US) images. They further showed how uncertainty estimation could be utilized to reject inferior-quality images and hence enhance the segmentation results.

Registration is a basic task in medical image analysis that can be used in numerous tasks including motion analysis, multimodal image alignment, intra-operative tracking and image segmentation. Zhu et al. [323] proposed a neural registration framework (NeurReg) with hybrid loss of displacement fields and data similarity, which considerably enhanced the existing SOTA methods of registration. They simulated different transformations with a registration simulator that created fixed-image and displacement field ground truth for training. They devised three segmentation frameworks based on the proposed registration framework: 1) MTL with atlas-based segmentation as an intermediate feature, 2) joint learning of both registration and segmentation tasks, and 3) atlas-based segmentation. Different probable ailments can be detected by accurate and automatic segmentation of anatomical structures in medical images. Bian et al. [324] introduced an uncertainty-aware domain alignment approach to address the domain shift issue in the cross-domain unsupervised domain adaptation (UDA) task. Domain shift is an issue related to the performance of the segmentation of various deep neural networks, and segmentation tasks may deteriorate several devices or modalities due to the notable dissimilarity across domains. In this regard, they specifically devised an uncertainty estimation and segmentation module (UESM) to attain the uncertainty map estimation. Then, they proposed a new uncertainty-aware cross entropy (UCE) loss to leverage the uncertainty information to enhance the performance of segmentation on extremely uncertain regions. The optimal target samples were selected through uncertainty guidance by an uncertainty-aware self- training (UST) method to further boost performance in the UDA task.

Kohl et al. [325] devised a segmentation network with a conditional variational autoencoder (cVAE), termed hierarchical probabilistic U-Net, that applied a hierarchical latent space decomposition. They demonstrated that their model formulation permitted reconstruction and sampling of segments with high fidelity while providing the flexibility to learn complex structured distributions across scales. Their model automatically split an inductive bias that they estimated to be useful in structured output prediction tasks beyond segmentation. In another work, Yin et al. [326] stated that uncertainty related to the fractional flow reserve (FFR) of coronary artery disease in a few properties such as anatomical and physiological properties is common. For this reason, they proposed a predictive probabilistic model for FFR using the BO approach. The outcomes clearly acknowledge the importance of dealing with uncertainty in the diagnosis of CAD. Li et al. [327] exploited uncertainty calibration within an AL framework for medical image segmentation. Uncertainty estimation is specifically crucial in a data-driven AL setting, where the goal is to attain definite accuracy with the least labeling effort. The model learns to choose the most enlightening unlabeled samples for annotation derived from its estimated uncertainty. Different acquisition strategies and uncertainty estimation techniques were explored. They argued that choosing regions to annotate instead of full images led to better-calibrated models.

It should be noted that UQ methods have been broadly used for medical image analysis. However, these methods have also been applied to analyze other types of medical data, including ECG [328], electrooculogram (EOG) [329], and EEG [330] signals as well as medical report information. Strodthoff et al. [328] investigated the performance of different CNNs for the prediction of age and gender to signal quality assessment considering uncertainty. In another study, Stoean et al. [329] used MC dropout with the DL method to automatically detect spinocerebellar ataxia type 2 in saccadic cases obtained from EOG data. The results showed the effectiveness of MC dropout for quantifying uncertainty. Moreover, Koulouri and

Rimpiläinen [330] proposed a model to quantify uncertainty using the Bayesian approximation error for focal source imaging and simultaneous skull conductivity. We provide further information about UQ methods applied in different medical application tasks in Table C.6 in the Appendix.

6.3. Natural Language Processing and Text Mining

In this part of the review, we briefly summarize some studies that have been conducted on UQ in the domain of NLP. It should be noted that we do not give the details of the methods due to page limitations. For this reason, we strongly recommend the reader to refer to the main sources if more information on the proposed UQ methods is needed. NLP focuses on understanding, analyzing and generating languages that humans utilize naturally [331]. In recent years, significant and practical real-world problems have been addressed, and large-scale systems have also been deployed in this research domain. Novel machine and deep learning approaches such as continuous space methods and DNNs have inferred language patterns from huge real-world data and have made accurate predictions about new data. One noteworthy challenge is to describe a language in a form that can be effectively processed by a learning system. NLP is an interdisciplinary field between linguistics and artificial intelligence [332]. One of the most broadly studied areas of NLP is text mining (TM), which collects vital information from free (unstructured) text. In this way, new knowledge can be extracted from a huge amount of text. However, the acquisition of reliable information from text is not straightforward because of the human linguistic ability to speak about nonexistent and unrealistic things or events. There are some propositions whose truth values cannot be unambiguously determined, as these propositions are uncertain, and they may be false in some possible worlds but true in others.

Uncertainty is a significant linguistic occurrence that is pertinent in many fields of language processing. In the most general case, it can be termed a lack of information, as the reader or listener is uncertain regarding a piece of information. Hence, uncertain propositions are those whose reliability or truth value cannot be determined due to a lack of information. Distinguishing between uncertain and factual (i.e., true and false) propositions is of prime importance in both natural language processing and linguistics applications. It is essential to recognize linguistic cues of uncertainty since the target and source language may differ in their frameworks of expressing uncertainty in machine translation. In clinical document classification, medical reports can be grouped depending on whether the patient probably suffers, does not suffer or suffers from an ailment. There are several different NLP applications that aim to investigate uncertainty in natural language texts in a couple of domains (e.g., news or biomedical texts). Most of these approaches use annotated databases for assessment. Various uncertainty corpora, such as the CoNLL-2010 Shared Task, FactBank, Genia and BioScope corpora, have been produced in recent years. Comparison of these corpora is not possible because of the lack of unified annotation principles. The prevailing uncertainty detectors are difficult to apply across domains, and novel resource creation for each domain is costly and time consuming. A high-dimensional hidden layer and a large dictionary size make training the RNN language model (RNN-LM) an ill-posed problem. Chien et al. [331] proposed a Bayesian approach to regularize the RNN-LM and utilized it for continuous speech recognition. The uncertainty of the estimated model parameters that was represented by a Gaussian prior was compensated for by penalizing the overly complicated RNN-LM. The Gaussian hyperparameter was estimated by maximizing the marginal likelihood, and regularized parameters were computed with reference to a maximum a posteriori criterion that was utilized to construct the regularized model. A small set of salient outer products were selected to derive the Bayesian RNN-LM (BRNN-LM) by developing a rapid approximation to a Hessian matrix. Clinical named entity recognition (NER) is one of the basic tasks for deriving clinical NLP systems. Domain experts are required for annotating a large amount of samples to achieve good performance with a machine learning (ML) system. This is an expensive exercise. A sample selection technique called AL aims to mitigate the annotation cost. Chen et al. [333] introduced and examined both novel and existing AL techniques for a clinical NER task to recognize medical treatments, problems and laboratory tests from clinical notes. They simulated AL experiments by applying different novel and prevailing algorithms in three categories, including baseline sampling, diversity-based, and uncertainty-based techniques. Based on the number of sentences vs. the learning curves of the F-measure, uncertainty sampling performed better than all its counterparts in terms of the area under the learning curve (ALC). Most diversity-based techniques yielded better performance than random sampling in ALC.

In another work, Kong et al. [334] introduced a novel theoretical perspective on data noising in RNN language models. They demonstrated that variants of data noising were instances of Bayesian RNNs with a particular variational distribution. They presented natural extensions of data noising under the variational

framework and a more principled method to apply during prediction by utilizing this insight. They devised an elementwise variational smoothing technique and variational smoothing with tied input and output embedding matrices. Their model was empirically tested on two language modeling datasets and exhibited superior performance to that of the prevailing data noising techniques. Factuality is a major concern in many domains, especially in social media, where informal texts are abundant. Existing methods in social media are dependent on lexical cues, where phrases are either omitted from a sentence or are expressed in substandard form. Han et al. [335] introduced ANFU, an attention-based neural framework for uncertainty identification on social media texts. ANFU incorporated a CNN to capture the most vital semantics and attention-based long short-term memory (LSTM) networks to represent the semantics of words. Experiments were performed on four benchmark datasets (2 English + 2 Chinese). Their proposed ANFU method performed better than any SOTA techniques in terms of the F1 score using four social media datasets. Zhang et al. [336] demonstrated that a huge deep learning model could utilize dropout VI to predict price movements from limit order books (LOBs), a data source with pricing and trading movements. To increase profit by avoiding needless trades and position sizing, uncertainty information extracted from posterior predictive distributions could be applied. Their experimental results showed that Bayesian techniques enhanced the predictive performance, as stochastic regularizers and uncertainty information could be utilized in trading. In another work, the authors of [337] designed a measure of uncertainty for long sequences of discrete random variables related to the words in the output sentence. This measure took account of epistemic uncertainty, similar to the MI applied for single discrete random variables such as in classification. Their uncertainty measures addressed a major intractability in the raw application of prevailing methods to long sentences. They utilized Europarl and WMT 13 for German-English translation tasks to train a Transformer model with dropout.

Machine translation is a popular topic in neural sequence-to-sequence models. A lack of diversity is observed in the final translations, and performance degradation is reported with large beams. The study [338] tried to uncover the extrinsic uncertainty caused by noisy training data and related this to some of the concerns associated with the inherent uncertainty of the task due to the existence of numerous valid translations for a single source sentence. They proposed metrics and tools to examine how uncertainty in the data was recorded by the model distribution and the effects of searching techniques in translation. They also presented tools for examining model calibration and some limitations of the current models that could be addressed by it. Accordingly, model calibration in classification was evaluated by Vaicenavicius et al. [339].

The authors of [340] designed a module for rapid experimentation with NN uncertainty and dubbed it Bayesian layers. NN libraries with drop-in replacements for common layers were extended by it. These layers recorded activations ("stochastic output layers"), pre-activation units (dropout), uncertainty overweight (BNNs) or the function itself (GP). They fused a 5-billion parameter "Bayesian Transformer" on 512 TPUv2 cores to determine uncertainty in a Bayesian dynamics and machine translation model for model-oriented planning. Bayesian layers could be utilized for probabilistic programming with stochastic processes used in the Edward2 language. On the other hand, the complexity of machine learning models poses uncontrolled risks, and the lack of control and knowledge of the internals of each component led to unavoidable effects, such as difficulty in auditability and a lack of transparency. Mena et al. [341] presented a wrapper that, given a black-box model, augmented its output prediction with an assessment of uncertainty. A decision rejection mechanism was employed to decrease the uncertainty or risk. They advocated for a rejection system based on the resulting uncertainty measure that discarded more uncertain predictions and selected more confident predictions; this improved the trustability of the system. They empirically showcased their method in simulated sentiment analysis framework for different domains.

Reliable UQ is a primary step towards devising accountable, transparent, and explainable artificial intelligence systems, and BDL plays a crucial role in such quantification. Xiao et al. [342] presented new strategies for examining data uncertainties and the benefits of characterizing models for NLP tasks. They utilized recurrent and CNN models to experiment empirically on language modeling, named entity recognition, and sentiment analysis to demonstrate that explicitly modeling uncertainties not only improved model performance but is also essential for computing output confidence levels in different NLP tasks. More studies have been conducted on the impact of Bayesian methods in improving the results of deep learning methods in NLP. The authors of [124] investigated a variational Bayes scheme for RNNs. First, they demonstrated that good-quality uncertainty estimates and superior regularization could be adapted by using truncated backpropagation with an additional computational cost during training and reducing the number of parameters by 80%. Second, they illustrated that the performance of Bayesian RNNs could be enhanced further

by employing a new kind of posterior approximation. The current batch statistics could be sharpened by incorporating local gradient information into the approximate posterior. This technique could be utilized broadly in training BNNs. They empirically showed that Bayesian RNNs performed better on an image captioning task and a language modeling benchmark than traditional RNNs. The authors of [343] proposed an intelligent framework to enhance en-route flight safety by trajectory prediction, where a Bayesian approach was utilized for model prediction uncertainty. Four steps were employed. In the first step, many raw messages were processed with a distributed computing engine, Apache Spark, to derive trajectory information efficiently. Two deep learning models were then trained to predict the flight trajectory from different perspectives. The deep learning models were blended together to create multi-fidelity prediction in the third step. Then, the multi-fidelity technique was expanded to multiple flights to examine safety based on the vertical and horizontal separation distance between two flights. The blended models showed promising results in en-route safety and flight trajectory prediction. Instead, a unified, widespread approach is needed that can be adapted to a particular need of each domain without much effort, and language independence in the model would also be preferable. Due to the large number of studies, we avoid explaining them in details. However, we provide a table in the Appendix that includes a summary of the most important UQ methods applied in the NLP domain (Table C.7).

6.4. Further Applications

In this section, we summarize more applications of various UQ methods. This section aims to cover a few of the most important recent studies.

As mentioned in previous sections, BDL deals with both epistemic and aleatoric uncertainty in predictions and has been successful in different domains such as climate change. Vandal et al. [344] devised a discrete-continuous BDL technique with lognormal and Gaussian likelihoods for uncertainty quantification. They presented a superresolution-based DL model dubbed "DeepSD" for statistical downscaling (SD) in climate that was utilized in predictions that followed a highly skewed distribution. Their discrete-continuous models performed better than a Gaussian distribution with respect to uncertainty calibration and predictive accuracy. In fact, traditional artificial neural networks (ANNs) lack the ability to model uncertainty and hence are not suitable for long-term planning tasks. Long-term ANN estimations deviate from the real behavior of the system due to approximation errors and process noise. In another work, Nalisnick et al. [345] presented two structured priors—automatic depth determination (ADD) and joint automatic relevance determination (ARD)-ADD—to permit Bayesian reasoning about a neural network's depth. The implementation led to higher runtime costs and little extra memory for BBB. Future work includes the use of structured variational approximations, comparison with other VI strategies and experiments on larger datasets.

New challenges arise in prevailing pixel-based prediction techniques with the advancement of remote sensing imagery. Although deep learning methods achieved a breakthrough in semantic segmentation of high-resolution images, most methods yield predictions with poor boundaries. Bischke et al. [346] proposed novel cascaded multitask loss for preserving semantic segmentation boundaries in satellite imagery. Their method outperformed the SOTA techniques by 8.3% without an additional postprocessing step. However, in autonomous driving, object detection plays a crucial role. Localizing objects and recognizing objects perfectly is infeasible due to incomplete data and sensor noise. Hence, the uncertainty associated with the predictions should be computed by the detector. Meyer et al. [347] devised a method that enhanced the learning of a probability distribution by taking into account potential noise in the ground-truth labeled data. Their method enhanced not only the object detection performance but also the accuracy of the learned distribution. RNNs have been applied to forecast increasingly complicated systems. Although the RNN literature is highly developed and expansive, UQ is often not taken into account. If it is considered, then the uncertainty is also usually quantified without the utilization of a rigorous approach. McDermott et al. [348] proposed a Bayesian RNN model for nonlinear spatiotemporal forecasting while quantifying uncertainty in a more formal framework. The unique nature of nonlinear spatiotemporal data was accommodated by modifying the basic RNN. They tested their model with two nonlinear spatiotemporal forecasting frameworks and a Lorenz simulation. On the other hand, RNN language models (RNNLMs) have proven their superiority in several different tasks including speech recognition. Learning the appropriate representation of contexts for word prediction can be achieved through the hidden layers of RNNLMs. Fixed hidden vectors and deterministic model parameters in conventional RNNLMs have limitations in modeling uncertainty over hidden representations. Yu et al. [349] presented a comparative study of hidden and parametric representation uncertainty

modeling techniques based on variational RNNLMs and Bayesian gates, respectively, that was examined on gated recurrent unit (GRU) and LSTM language models. Performance improvements were observed over conventional RNNLMs with their model in terms of the word error rate and perplexity.

Predictive accuracy in black-box turbulence models is enhanced by tuning Reynolds-averaged Stokes (RANS) simulations and applying machine learning algorithms. Geneva et al. [350] presented a new data-driven approach to provide probabilistic bounds for fluid quantities and enhanced RANS predictions. The anisotropic tensor component of Reynolds stress was predicted by using an invariant BDNN. The Stein variational gradient descent algorithm was applied to train the model. Based on the proposed method, the associated probabilistic bounds and prediction enhancement of the data-driven model were addressed. Following the research for dealing with uncertainty, we came across a study of Feng et al. [351] that proposed a novel extreme learning machine (ELM) termed rough ELM (RELM). RELM utilized rough sets to divide data into a lower approximation set and upper approximation set, and these sets were used to train lower approximation neurons and upper approximation neurons. RELM showed a comparable accuracy and repeatability in most classification tasks. In another study, Walmsley et al. [352] applied a Bayesian CNN and a new generative model of Galaxy Zoo volunteer responses to infer posteriors for the visual morphology of galaxies. The probability of each possible label could be predicted by using a Bayesian CNN to learn from galaxy images with uncertain labels. Their posteriors were reliable for practical use, as they were well calibrated. They utilized a BALD AL strategy applying their posteriors to request volunteer responses for a subset of galaxies. They demonstrated that training their Bayesian CNNs utilizing AL, they needed up to 35-60% fewer labeled galaxies, instead relying on morphological features.

The distribution of states at execution time may differ from the distribution observed during training, and this makes learning a policy utilizing only observational data a challenging task. Henaff et al. [353] introduced the idea of training a policy by unrolling a learned model of environment dynamics over multiple time steps while explicitly penalizing on the basis of two costs: the original cost the policy sought to optimize and an uncertainty cost that represented its divergence from the states it was trained on. They examined their strategy utilizing a huge observational dataset of driving behavior recorded from traffic cameras. In drug discovery, as another application of UQ methods, it is a challenge to predict the physical properties and bioactivity of small molecules. Zhang et al. [354] used Bayesian semi-supervised graph convolutional neural networks to achieve UQ and AL. Sampling from a posterior distribution was applied in this Bayesian approach, which estimates uncertainty in a statistically principled way. Semi-supervised learning untangled regression and representation learning, allowing the model to start AL from small training data and keeping uncertainty estimates accurate with a low data limit. Their method highlighted the promise of BDL in chemistry.

According to the literature, it is obvious that machine learning has the potential to give valuable assistance in clinical decision making, especially in the intensive care unit (ICU) of a hospital. Traditional machine learning models do not take into account uncertainty in predictions. Ruhe et al. [355] showed how predictive uncertainty and Bayesian modeling could be utilized to recognize out-of-domain examples and reduce the risk of faulty predictions. They utilized BNNs to predict the risk of mortality of ICU patients. Their empirical results showed that uncertainty could be used to detect out-of-domain patients and avoid probable errors. Many machine learning techniques need human supervision to yield optimal performance. The quality of manual annotations is essentially limited in methods such as DensePose. Neverova et al. [356] addressed the issue by augmenting neural network predictors with the ability to output a distribution over labels, thus introspectively and explicitly capturing the aleatoric uncertainty in the annotations. New SOTA accuracy on the benchmark could be achieved by having a better understanding of uncertainty and hence solving the original DensePose task more accurately. The uncertainty estimates produced by multiple models could be used in fusing predictions in a better way to ensemble models that could enhance accuracy further.

As mentioned earlier, uncertainty estimates in RL tasks and large vision models can be obtained via dropout. A grid search over the dropout probabilities is essential — this is impossible with RL and a prohibitive operation with large models. Gal et al. [357] devised a novel dropout variant that attained better performance and improved the calibrated uncertainties. They used a continuous relaxation of dropout's discrete masks that was based on recent advancements in BDL. They analyzed their variant on several tasks and provided insights into usual practice in areas where larger dropout probabilities are often utilized in deeper model layers.

Mobile robots for indoor use depend on 2D laser scanners for navigation, localization and mapping.

These sensors are unable to measure the full occupancy of complex objects and cannot detect transparent surfaces. These estimates are prone to uncertainty, making the evaluation of confidence a significant issue for autonomous navigation and mapping. Verdoja et al. [358] proposed a solution to the problem using CNNs, as another application of UQ methods. They demonstrated that uncertainty regarding obstacle distances was, however, better modeled with a Laplacian distribution. They created maps based on DNN uncertainty models. Their algorithm was used to create a map that included information on obstacle distance estimates while taking account of the level of uncertainty in each estimate. Traditional high-dimensional data reduction methods such as projection pursuit regression (PPR), reduced rank regression (RRR), partial least squares (PLS), and principal component analysis (PCA) are all shallow learners. Polson et al. [359] examined DL counterparts that exploited multiple deep layers of data reduction and provided predictive performance gains. Dropout regularization and SGD training optimization provided variable selection and estimation. They illustrated their technique by providing an analysis of international bookings on Airbnb.

Energy (e.g., electricity markets) is another common domain for the application of different UQ methods with machine and deep learning. Successful participation in liberalized electricity markets can be achieved by forecasting accurate day-ahead energy prices. Brusaferri et al. [360] presented a new approach based on BDL strategies for probabilistic energy price forecast. Scalability to a complex network was guaranteed by executing a specific training method. They examined their system on two day-ahead markets characterized by different behaviors. The robust performance of the system was achieved by providing forecast uncertainty indications in out-of-sample conditions. Moreover, class imbalance in remote sensing poses a challenge for land cover mapping, where small objects get less attention in yielding better accuracy. Uncertainty quantification on pixel levels using CNNs is another issue in remote sensing. Kampffmeyer et al. [361] devised a deep CNN in remote sensing images for land cover mapping primarily aiming at urban areas. Their method aimed to achieve good accuracy for small objects. They applied recent technologies for UQ in the domain of remote sensing. This method yielded an overall classification accuracy of 87%. On the other hand, accurately predicting the net load arising from distributed photovoltaic (PV) generation is a great challenge. Sun et al. [362] presented a new probabilistic day-ahead net load forecasting technique to capture both aleatoric uncertainty and epistemic uncertainty utilizing BDL. The performance of aggregated net load forecasting was improved by considering residential rooftop PV outputs as the input features, and it exploited clustering in subprofiles. The proposed scheme proved its efficacy with high PV visibility and subprofile clustering.

Different UQ methods have also been applied to recommender systems, or recommendation systems. For example, Zeldes et al. [363] proposed a new recommendation system named Deep Density Networks (DDN), which combined content-based DL methods with a collaborative scheme to estimate uncertainty and determine a robust model. Sun et al. [364] introduced a multi-aspect user interest model based on uncertainty theory and sentiment analysis for recommendation systems, called MAUI (*multi-aspect user-interest*). The proposed model was tested on five products obtained from three e-commerce websites. Uncertainty theory was used to estimate the similarity of all user opinions. Moreover, Jasberg and Sizov [365] showed that Bayesian brain theory can use the same models as Bayesian recommender systems. The impacts of these theories are discussed in cases such as neural noise, decision making, and determining user preferences. The obtained outcomes support integrating cognitive neuroscience into different intelligent systems to further improve the prediction of a wide variety of human behaviors.

7. Discussion

We reviewed most UQ studies on machine and deep learning methods and briefly discussed their methods in the previous sections. In what follows, we discuss the main advantages and disadvantages of the main UQ methods. Moreover, we discuss several important literature gaps and open issues blocking the development of UQ methods. Now, we briefly discuss the main advantages and disadvantages of a few UQ methods. Table 4 lists the most important advantages and disadvantages of several well-known UQ methods in the literature.

7.1. Literature Gaps and Open Issues

In what follows, we list the most important gaps and open issues that should be addressed in the near future (see 7.1.1). Furthermore, we list a few future research directions for further studies, and the most

Table 4: Advantages and disadvantages of UQ methods.

Method		Advantage	Disadvantage
Bayesian	MC	(1) No need to change the model training process, (2) Low training complexity, (3) Easy to implement.	(1) Not very reliable for OoD data, (2) Needs multiple samplings during inference.
	MCMC	(1) Computationally more intensive compared to VI, (2) Asymptotically guarantees of producing exact samples.	(1) Very slow, (2) Fail to find poor convergence, (3) High MC error.
	VI	(1) Very fast (faster than MCMC), (2) Benefiting from stochastic optimization methods, (3) Suited to big datasets.	(1) Heavily depend on the starting point, (2) Very complicated calculations.
	BAL	(1) Able to learn from small amounts of data, (2) Able to add samples with high classification uncertainty to training.	(1) Lack of scalability to high-dimensional data, (2) Difficult to quantify loss functions.
	BBB	(1) Returning the posterior over the weights, (2) Allowing more complicated prior distributions.	(1) Requiring extra sweep over KL trade-off coefficients, (2) More parameters to train (approximately two times).
	VAE	(1) Easy to optimize its loss, (2) Mapping an input sample in the original data to latent factors.	(1) Collapse in latent space, (2) Difficult to interpret the code, (3) Low quality of the generated sample images.
Ensemble	DE	(1) Robust prediction, (2) Can be considered as base learners, (3) Limiting the dispensable sensitivity of particular training data, (4) Robust uncertainty estimates.	(1) More resource consuming, (2) Time consuming, (3) Weak performance on smaller problems.
	DEB or BDE	(1) Can perform batter than DEs in OoD settings, (2) Emulating the analytic posterior predictive.	(1) Weaker than standard DEs in not detrimental confident predictions, (2) Lazy learning procedure.

important gaps and open issues as well as future application directions (see 7.1.2) are discussed below. Finally, we briefly discuss the lack of data and code availability in the literature, particularly in the application of UQ methods (see 7.2).

7.1.1. General Gaps

In this subsection, the main general research gaps and open issues are discussed. The most important literature gaps and open issues in the application of UQ methods are listed.

- This review shows that most of the proposed UQ methods are presented for supervised learning methods, followed by unsupervised. However, we note that there are fewer studies on semisupervised learning methods. We believe that this is an important gap in the domain of UQ that could be filled in the future.

- Our findings reveal that most UQ methods by far have been proposed for different types of NNs, especially DL methods. However, there are many other methods in the field of ML for which uncertainty has either not been investigated or has been superficially discussed. The probable reason for this is that DL methods have been among the best (SOTA) methods in various fields (e.g., computer vision, medical image analysis, NLP, time series analysis, signal processing). However, as a matter of fact, it can be claimed that different types of traditional ML methods have significantly good performance in the analysis of small data, whereas DL is almost incapable of performing this task.

- Fusion-based methods (e.g., multimodal Bayesian fusion [95], multilevel fusion [366], image fusion [367], data fusion [368]) have shown great ability to optimize different machine and deep learning methods. This led us to investigate the effects of a variety of fusion-based methods in dealing with uncertainty in machine and deep learning methods. We realized that fusion-based methods have significant potential to address the uncertainty of models. Therefore, we suggest that more fusion-based methods could be used in future work for quantifying uncertainty.

- The results of our research show that although a variety of ensemble methods [369, 370, 371, 372, 373, 374, 375] also have a great ability to deal with uncertainty along with good performance and ability to optimize the performance of other methods, the good capabilities of these methods have not been used more widely. In other words, we noticed that these methods have performed remarkably well in few studies. However, we realized that the ensemble methods are less commonly used in recent studies. Therefore, we strongly recommend further studies on ensemble methods and their substantial impact on quantifying uncertainty in the machine and deep learning domain. For example, Caldeira and Nord [376] presented a comparative study to compare the performance of three important UQ methods: BNN, CD and DE. Based on the obtained outcomes, they recommended DE for further investigation and applications, as it achieved the best or comparable results in the study.

- Decision making is a cognitive process that results in choosing the best possible action or belief among all other alternative options. There are few well-known theories, such as three-way decisions [377] and info-gap (IG) decisions [378, 379], that can be used as UQ methods. Our findings reveal that these theories have been able to help significantly in dealing with uncertainty. For this reason, we believe that various decision making theories can be used during the decision making process for machine and deep learning methods.

- Active learning (AL, sometimes also called optimal experimental design) plays a key role in dealing with a lack of labeled data to label new data with the desirable outputs, as most researchers in the domain of machine and deep learning are aware that obtaining labeled data is very difficult, costly and time consuming. In a few extremely sensitive areas such as healthcare and self-driving cars, this importance becomes more apparent. We have reviewed several good studies in this area, including Gordon et al. [380], Lee et al. [381], Nguyen et al. [382], Hu et al. [383] and Qu et al. [384]. Our reviews, however, reveal that although uncertainty in this area is quite important, very few studies have been done on this subject. For example, Sinha et al. [385] proposed a new adversarial Al called Variational Adversarial AL (VAAL). The proposed model trained a latent space by using a VAE and an adversarial network trained to discriminate between labeled and unlabeled data (see Fig. 22). They

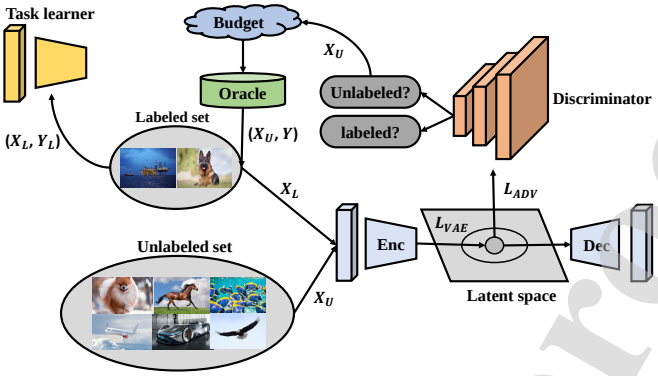


Fig. 22: A schematic view of the VAAL model which is reproduced based on [385].

showed the dramatic impact of this type of method on UQ. For this reason, we believe that researchers can fill this gap with further studies in order to improve the data labeling quality by obtaining far more certainty than previous studies.

- Transfer learning is a technique to deal with training different machine and deep learning methods well when there is not enough data to train the models properly. In other words, transfer learning is a technique for repurposing and reusing already trained machine and deep learning methods in a set of new situations. Based on our review study, this technique also has uncertainty. Hence, we would recommend conducting more research in this area and proposing new UQ methods for transfer learning.
- Neural architecture search (NAS) methods are a set of techniques for automatically designing ANNs. Uncertainty awareness in such methods is an important and sensitive area for the use of UQ methods. However, the results of our research reveal that very few studies on this issue have been done so far. Hence, we list this case as another research gap for further investigation. Moreover, neural ensemble search methods [386] are well-developed techniques for uncertainty calibration.
- Self supervised learning (SSL) [387, 388] is an important subset of unsupervised learning for generating output labels from different data objects by finding a relationship between various parts of the object or a variety of views of the given object. We think that SSL methods have several sources of uncertainty, and therefore, further investigation of these methods has a high potential as an important research gap of UQ.
- The attention mechanism [389, 390] is a powerful strategy in NNs and DL methods to concentrate more on adequate parts of the input data than on unnecessary (irrelevant) parts while performing a prediction task. However, we found that selecting relevant and irrelevant parts of the data is accompanied by uncertainty. Our reviews show that fewer studies of UQ have been conducted in this area. For this reason, we also list this area as a research gap for further investigation.
- OoD [391, 392, 393, 394, 395, 396] inputs (samples) can improve the robustness of different models as well as increase the uncertainty. For example, Lee et al. [381] proposed a new model called BTAML for dealing with imbalanced data and the detection of OoD samples. According to previous studies, we can see that detection of OoD samples can help in achieving outstanding performance with different NN and DL methods. However, quantifying uncertainty in detecting OoD samples needs further investigation [397].
- Hypernetworks [398, 399, 400] are a very powerful method of generating weights for one network using another network. We found that quantifying uncertainty in hypernetworks can be a very useful approach to achieving a better defense in dealing with adversarial samples. However, we found very few studies dealing with uncertainties in hypernetworks [401]. Hence, we can also suggest this as an open issue on which more research can be conducted to deal with uncertainty in hypernetworks.

- Continual learning (CL) [402, 403, 404, 405, 406, 407] (or continual deep learning [408]) is a subset of machine learning that provides an ability for different models to continually learn from a stream of data. In other words, an algorithm deals with a sequence of samples (of stationary data) while reusing former knowledge and then exploits it to better adapt to a changeable environment. Our review shows that there are only a few studies introducing UQ methods in CL. Therefore, quantifying uncertainties in the CL domain is another open research field for applying new UQ methods of dealing with uncertainties.

- Graph neural networks (GNNs) [409, 410, 411, 412, 413] are powerful graph representation learning methods that extract high-level features of related nodes from their topological neighborhoods. Uncertainty quantification for GNNs in graph analytic tasks is still an open challenge. Our findings reveal that even though Bayesian-based GNNs and GCNs (graph convolutional networks) [414, 415, 416, 417] and deep graph BO [418] have shown promising outcomes for quantifying uncertainties in GNNs and GCNs, there are a limited number of studies in this research domain. Hence, we recommend proposing new efficient UQ methods for dealing with uncertainties in GNNs and GCNs.

- BO [419, 420] is a sample-efficient global optimization method for optimizing time-consuming black-box objective functions, which take quite a long time to evaluate. We found that BO can be a very efficient approach not only for optimizing a wide range of applications including hyperparameter tuning [421] but also for quantifying uncertainties in machine and deep learning methods and quantifying uncertainties in BO; this remains as an open issue for future investigation [422, 423].

- Uncertainty calibration is another approach for measuring a model's calibration errors that has been used in different case studies [169, 424, 425, 426]. However, we noticed that there are fewer studies on the calibration of modern NNs, DLs and many other ML methods [427, 425, 428, 429, 430, 431]. Calibration measures can then be used for quantifying uncertainties in ML and DL predictions. Hence, we recommend developing different uncertainty calibration approaches.

7.1.2. Future Directions Based on Applications

UQ methods have been making highly regarded achievements and obtaining distinguished performance when applied to different machine and deep learning methods. However, there are some open challenges that should be addressed. We provide several future directions of UQ methods for three main research topics: computer vision and image processing, medical image processing and NLP.

Computer Vision and Image Processing: As discussed earlier, computer vision and image processing are two main research domains for the application of different UQ methods. Although various studies have been conducted in these areas, there are certainly still many open research directions that should be considered in future studies. In the following, we aim to list a few of the most important future directions in these domains.

Theoretical analysis and a more resilient inference methodology for various UQ methods should be investigated in future studies. For example, the integration of semisupervised learning and AL can be developed for acquiring new samples. In addition, data labeling is a time-consuming and costly process in all domains, not only computer vision and image processing. Therefore, we recommend conducting further studies on automated data labeling techniques and investigating the impact of UQ methods. Applications of cascade structures have proven to be a powerful mechanism for improving various machine and deep learning methods. However, we think that simplifying these methods and their integration with UQ methods for different computer vision and image processing tasks is valuable. Moreover, the integration of dynamic and multimodal image restoration issues with some advanced inversion approaches (e.g., different plug-and-play schemes) for applying UQ methods to reveal relevant point estimates is another interesting future research direction. In addition, the review outcomes reveal that ensemble methods are still among the best approaches, especially for detecting epistemic uncertainties (OoD issues). For this reason, the application of new ensemble methods is another interesting research direction in computer vision and image processing.

The integration of UQ methods with different human pose architectures followed by using the estimated uncertainty in future frame prediction and motion tracking tasks is another engaging open research direction. Additionally, although we mentioned above some UQ methods for BAL, we note that better uncertainty estimates in BAL as well as more complex methods should be proposed in this domain. In addition, we found

that detecting adversarial samples is an interesting challenge for BNNs [432]. Thus, we strongly recommend further studies to develop more effective UQ methods for detecting adversarial samples. Sampling-free learning methods (e.g., Bayesian quantized networks (BQNs) [433], sampling-free VI [434]) are powerful techniques for learning an appropriate posterior distribution over their discrete parameters with truly calibrated uncertainty. Furthermore, embedding techniques have obtained outstanding performance in different computer vision and image processing tasks. However, we found that there are very few studies on probabilistic embedding strategies [435, 436] that quantify uncertainty. We also note that even though Bayesian methods (i.e., variational Bayes [437]) have been used for UQ in capsule routing methods, the calibration of uncertainty estimates of predictions by using different capsule networks is an open future research direction. Online applications of different BNNs are an open issue for future investigations for various reasons such as the limitations of variational techniques and risks in selecting the appropriate approximations of batch posteriors [438]. The uncertainty of CL [439] is another open research direction in computer vision and image processing. For example, Nguyen et al. [440, 441] proposed variational CL (VCL) to deal with uncertainty and showed the effectiveness of such an approach. Finally, we found that quantifying uncertainty in multitask transfer learning [442] is a very important research domain in which further investigations are highly warranted.

Medical Image Analysis: One possible research direction that could be considered in the future is a closer collaboration between medical and artificial intelligence researchers. Due to the high sensitivity in this field of science, we strongly recommend collecting larger medical data in this domain. This can be very helpful in resolving uncertainty, and as a result, the proposed machine and deep learning methods can perform better in predicting various diseases and cancers. Ground-truth data for medical image segmentation plays a critical role in the correctness of the obtained results. For this reason, closer cooperation between the two groups can provide platforms for optimizing existing machine and deep learning models. Furthermore, the referral of incorrectly predicted data to specialists has a great role in dealing with uncertainty. Hence, there is a need for close collaboration between medical and computer researchers in the field of medical image segmentation.

We also note that various fusion methods have good potential in segmentation uncertainty prediction for medical images. Moreover, we found that in most previous studies, standard losses for medical segmentation were used, whereas there are some new but effective losses that could be used. On the other hand, a combination of both visualization of uncertainty and medical image segmentation can be used in AL computer assisted diagnostic (CAD) systems to improve the segmentation output. Another important future direction could focus on big medical data collection. Having more data can dramatically improve the performance of various deep and machine learning methods. Our comprehensive review reveals that the problem of having sufficient medical data is still open. However, if this is not possible, transfer learning techniques can be an ideal solution for improving the training process. Using this technique, we can properly tune the applied DL methods; however, we know there are a few uncertainties. As mentioned above, this can be considered an open gap for future researchers.

The development of different semisupervised methods for medical image segmentation is another promising approach for dealing with medical data shortages. We found that UQ methods can have great impact in semisupervised medical image segmentation, which is another interesting research direction for future investigations. Along with all of these open directions in medical data analysis, MTL has also shown promising performance for medical data analysis. However, as Nguyen et al. [443] showed, adding UQ methods to MTL can significantly quantify uncertainty for the prediction of clinical risks. However, as Gast and Roth [283] stated, although probabilistic approaches have been used widely for several years, probabilistic approaches have not been comprehensively applied in practice since sampling techniques are frequently too slow. To address this issue, the development of proper probabilistic approaches [283] can be used in real medical applications.

Natural Language Processing (NLP): There are a few NLP-based subjects such as neural machine translation [444] and some other interdisciplinary subjects such as image captioning [445, 446, 447, 448] and visual question answering (VQA) [449, 450, 451, 452, 453] that are closely associated with NLP. Finding the right caption for an image has always been a challenge in computer science. In particular, this issue can be accompanied by some uncertainties due to the merging of two important disciplines (i.e., image processing and NLP). On the other hand, medical VQA is a very important task for health organizations to determine the most appropriate answers to the given questions. Indeed, this topic involves both image processing and

NLP tasks at the same time. We believe that because of the essence of this matter, adding methods to deal with uncertainty can greatly contribute to the productivity of this branch of research. In addition, classification of data stream text environments is an interesting domain for the application of UQ methods for finding uncertain sentences.

Further Directions: In this section, we discuss a few research directions in various subjects such as RL, numerical data analysis, signal processing and toy and synthetic data. For instance, the application of meta RL (MRL) is effective and efficient in optimizing the decision performance. However, the decision-making process always comes with uncertainty. Hence, we suggest adding UQ methods with various MRL models to yield better decisions with more certainty. The proposed natural gradient approach [253] can be generalized with some other approximation types (e.g., exponential-family distributions) and applied to RL and stochastic optimization. Moreover, achieving a proper understanding of the interaction between choosing an inference algorithm and an approximating family is another future research direction for synthetic data (a regression task). Developing various stochastic regularization methods is another open direction for researchers. We also note that leveraging the proper weights of the Frank-Wolfe optimization algorithm [117] and determining how this technique interacts with some alternative procedures of approximate inference can be interesting avenues for further investigation.

Moreover, digital healthcare is a very important research area that can help to make medicine more precise and personalized. Quantifying uncertainty in digital healthcare and deploying it in real-world clinical settings is another open research path. Approximate Bayesian inference in CL and sequential decision-making applications should be used as an inner procedure of a larger method. In addition, this procedure needs a robust version of BNNs. Hence, the application of deterministic VI [454] with different BNNs is an ideal approach. Accordingly, the learning of different BNNs [455, 254] can be optimized by using various assumed density filtering (ADF) techniques [456] and applying them in different machine and deep learning tasks (e.g., NLP, computer vision and image processing, signal processing). In addition, ensemble-based sampling [457, 458] methods have shown the ability to approximate sampling techniques (i.e., Thompson sampling) and properly deal with uncertainty in complex methods such as different NNs. Finally, quantifying uncertainties for multi-agent systems [459, 460] is another important future direction since an individual agent cannot solve problems that are impossible or difficult.

7.2. Lack of Data and Code Availability

The availability of data and code plays a significant role in improving prior methods. In other words, having data and code will assist researchers in examining proposed methods and finding the main gaps. Our comprehensive review shows that most previous studies (especially medical case studies) do not share their data and code with others. We understand that because there are few cases, the authors may not be able to make the data or code public, but we believe sharing both the data and code will be very helpful in improving the quality and performance of different machine and deep learning methods. In other words, having the code and data of a research paper will accelerate the optimization of the code in the study.

8. Conclusion

Uncertainty quantification (UQ) is one of the key parts of the decision-making process. UQ methods have become popular for evaluating uncertainty in various real-life applications, being an inseparable part of traditional machine learning and deep learning techniques. This study provides a comprehensively review of popular and efficient UQ methods that have been applied in traditional machine learning and deep learning. Precisely, we presented a comprehensive description and comparative analysis of current state-of-the-art UQ approaches. Moreover, we reviewed several UQ perspectives, including common UQ applications, in-depth advantage discussions and main research gaps, ending by presenting some solid future research directions in this area. We believe that this review paper focusing on the use of UQ methods in artificial intelligence will benefit researchers in a variety of fields and may be considered a guideline for the use of UQ in deep learning-based applications.

Acknowledgment

This work was partially supported by the Australian Research Council's Discovery Projects funding scheme (project DP190102181) and the Natural Sciences and Engineering Research Council of Canada (grant no. 249644).

References

[1] H. Jiang, B. Kim, M. Guan, M. Gupta, To trust or not to trust a classifier, in: Advances in neural information processing systems, 2018, pp. 5541–5552.

[2] A. Malinin, Uncertainty estimation in deep learning with application to spoken language assessment, Ph.D. thesis, University of Cambridge (2019).

[3] T. M. Mitchell, The need for biases in learning generalizations, Department of Computer Science, Laboratory for Computer Science Research ..., 1980.

[4] V.-L. Nguyen, S. Destercke, E. Hüllermeier, Epistemic uncertainty sampling, in: International Conference on Discovery Science, Springer, 2019, pp. 72–86.

[5] C. C. Aggarwal, X. Kong, Q. Gu, J. Han, S. Y. Philip, Active learning: A survey, in: Data Classification, Chapman and Hall/CRC, 2014, pp. 599–634.

[6] B. T. Phan, Bayesian deep learning and uncertainty in computer vision, Master's thesis, University of Waterloo (2019).

[7] Y. Gal, Uncertainty in deep learning, Ph.D. thesis, University of Cambridge (2016).

[8] E. Hüllermeier, W. Waegeman, Aleatoric and epistemic uncertainty in machine learning: A tutorial introduction, arXiv preprint arXiv:1910.09457 (2019).

[9] C. Hubschneider, R. Hutmacher, J. M. Zöllner, Calibrating uncertainty models for steering angle estimation, in: IEEE Intelligent Transportation Systems Conference, 2019, pp. 1511–1518.

[10] J. Mukhoti, Y. Gal, Evaluating bayesian deep learning methods for semantic segmentation, arXiv preprint arXiv:1811.12709 (2018).

[11] H. Kabir, M. Abdar, S. M. J. Jalali, A. Khosravi, A. F. Atiya, S. Nahavandi, D. Srinivasan, Spinalnet: Deep neural network with gradual input, arXiv preprint arXiv:2007.03347 (2020).

[12] X. Wang, Y. Zhao, F. Pourpanah, Recent advances in deep learning, International Journal of Machine Learning and Cybernetics (2020).

[13] E. Begoli, T. Bhattacharya, D. Kusnezov, The need for uncertainty quantification in machine-assisted medical decision making, Nature Machine Intelligence 1 (1) (2019) 20–23.

[14] D. Hafner, D. Tran, T. Lillicrap, A. Irpan, J. Davidson, Noise contrastive priors for functional uncertainty, arXiv preprint arXiv:1807.09289 (2018).

[15] C. Ning, F. You, Optimization under uncertainty in the era of big data and deep learning: When machine learning meets mathematical programming, Computers & Chemical Engineering 125 (2019) 434–448.

[16] H. D. Kabir, A. Khosravi, M. A. Hosen, S. Nahavandi, Neural network-based uncertainty quantification: A survey of methodologies and applications, IEEE access 6 (2018) 36218–36234.

[17] D. E. Rumelhart, G. E. Hinton, R. J. Williams, Learning internal representations by error propagation, Tech. rep., California Univ San Diego La Jolla Inst for Cognitive Science (1985).

[18] A. D. Kiureghian, O. Ditlevsen, Aleatory or epistemic? does it matter?, Structural Safety 31 (2) (2009) 105 – 112, risk Acceptance and Risk Communication.

[19] A. Kendall, Y. Gal, What uncertainties do we need in bayesian deep learning for computer vision?, in: Advances in neural information processing systems, 2017, pp. 5574–5584.

[20] S. Kullback, R. A. Leibler, On information and sufficiency, The annals of mathematical statistics 22 (1) (1951) 79–86.

[21] D. M. Blei, A. Kucukelbir, J. D. McAuliffe, Variational inference: A review for statisticians, Journal of the American statistical Association 112 (518) (2017) 859–877.

[22] Y. Gal, Z. Ghahramani, Bayesian convolutional neural networks with bernoulli approximate variational inference, arXiv preprint arXiv:1506.02158 (2015).

[23] M. I. Jordan, Z. Ghahramani, T. S. Jaakkola, L. K. Saul, An introduction to variational methods for graphical models, Machine learning 37 (2) (1999) 183–233.

[24] K.-C. Wang, P. Vicol, J. Lucas, L. Gu, R. Grosse, R. Zemel, Adversarial distillation of bayesian neural network posteriors, arXiv preprint arXiv:1806.10317 (2018).

[25] L. V. Jospin, W. Buntine, F. Boussaid, H. Laga, M. Bennamoun, Hands-on bayesian neural networks–a tutorial for deep learning users, arXiv preprint arXiv:2007.06823 (2020).

[26] H. Wang, D.-Y. Yeung, Towards bayesian deep learning: A survey, arXiv preprint arXiv:1604.01662 (2016).

[27] J. Maroñas, R. Paredes, D. Ramos, Calibration of deep probabilistic models with decoupled bayesian neural networks, Neurocomputing (2020).

[28] P. Izmailov, W. J. Maddox, P. Kirichenko, T. Garipov, D. Vetrov, A. G. Wilson, Subspace inference for bayesian deep learning, in: Uncertainty in Artificial Intelligence, PMLR, 2020, pp. 1169–1179.

[29] A. D. Cobb, B. Jalaian, Scaling hamiltonian monte carlo inference for bayesian neural networks with symmetric splitting, arXiv preprint arXiv:2010.06772 (2020).

[30] T. Karaletsos, T. D. Bui, Hierarchical gaussian process priors for bayesian neural network weights, arXiv preprint arXiv:2002.04033 (2020).

[31] A. Foong, D. Burt, Y. Li, R. Turner, On the expressiveness of approximate inference in bayesian neural networks, Advances in Neural Information Processing Systems 33 (2020).

[32] A. Kucukelbir, D. Tran, R. Ranganath, A. Gelman, D. M. Blei, Automatic differentiation variational inference, J. Mach. Learn. Res. 18 (1) (2017) 430–474.

[33] R. M. Neal, Bayesian learning for neural networks, Vol. 118, Springer Science & Business Media, 2012.

[34] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, R. Salakhutdinov, Dropout: A simple way to prevent neural networks from overfitting, Journal of Machine Learning Research 15 (56) (2014) 1929–1958.

[35] Y. Gal, Z. Ghahramani, Dropout as a bayesian approximation: Representing model uncertainty in deep learning, in: international conference on machine learning, 2016, pp. 1050–1059.

[36] P. McClure, N. Kriegeskorte, Representing inferential uncertainty in deep neural networks through sampling, in: International Conference on Learning Representations, ICLR 2017-Conference Track Proceedings, 2016.

[37] K. Brach, B. Sick, O. Dürr, Single shot mc dropout approximation, arXiv preprint arXiv:2007.03293 (2020).

[38] G. Wang, W. Li, M. Aertsen, J. Deprest, S. Ourselin, T. Vercauteren, Aleatoric uncertainty estimation with test-time augmentation for medical image segmentation with convolutional neural networks, Neurocomputing 338 (2019) 34 – 45.

[39] H. Liu, R. Ji, J. Li, B. Zhang, Y. Gao, Y. Wu, F. Huang, Universal adversarial perturbation via prior driven uncertainty approximation, in: Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 2941–2949.

[40] T. Nair, D. Precup, D. L. Arnold, T. Arbel, Exploring uncertainty measures in deep networks for multiple sclerosis lesion detection and segmentation, Medical Image Analysis 59 (2020) 101557.

[41] A. Amini, A. Soleimany, S. Karaman, D. Rus, Spatial uncertainty sampling for end-to-end control, arXiv:1805.04829 (2018).

[42] J. Tompson, R. Goroshin, A. Jain, Y. LeCun, C. Bregler, Efficient object localization using convolutional networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 648–656.

[43] L. Yu, S. Wang, X. Li, C.-W. Fu, P.-A. Heng, Uncertainty-aware self-ensembling model for semi-supervised 3d left atrium segmentation, in: D. Shen, T. Liu, T. M. Peters, L. H. Staib, C. Essert, S. Zhou, P.-T. Yap, A. Khan (Eds.), Medical Image Computing and Computer Assisted Intervention, Springer International Publishing, Cham, 2019, pp. 605–613.

[44] A. Kendall, V. Badrinarayanan, R. Cipolla, Bayesian segnet: Model uncertainty in deep convolutional encoder-decoder architectures for scene understanding, arXiv preprint arXiv:1511.02680 (2015).

[45] V. Badrinarayanan, A. Kendall, R. Cipolla, Segnet: A deep convolutional encoder-decoder architecture for image segmentation, IEEE Transactions on Pattern Analysis and Machine Intelligence 39 (12) (2017) 2481–2495.

[46] C. Leibig, V. Allken, M. S. Ayhan, P. Berens, S. Wahl, Leveraging uncertainty information from deep neural networks for disease detection, Scientific reports 7 (1) (2017) 17816.

[47] S. Choi, K. Lee, S. Lim, S. Oh, Uncertainty-aware learning from demonstration using mixture density networks with sampling-free variance modeling (2017). arXiv:1709.02249.

[48] C. M. Bishop, Mixture density networks, Tech. rep., Neural Network Research Group (1994).

[49] A. Jungo, R. McKinley, R. Meier, U. Knecht, L. Vera, J. Pérez-Beteta, D. Molina-García, V. M. Pérez-García, R. Wiest, M. Reyes, Towards uncertainty-assisted brain tumor segmentation and survival prediction, in: A. Crimi, S. Bakas, H. Kuijf, B. Menze, M. Reyes (Eds.), Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries, Springer International Publishing, Cham, 2018, pp. 474–485.

[50] T. Pohlen, A. Hermans, M. Mathias, B. Leibe, Full-resolution residual networks for semantic segmentation in street scenes (2016). arXiv:1611.08323.

[51] K. Wickstrøm, M. Kampffmeyer, R. Jenssen, Uncertainty modeling and interpretability in convolutional neural networks for polyp segmentation, in: IEEE International Workshop on Machine Learning for Signal Processing (MLSP), 2018, pp. 1–6.

[52] E. Shelhamer, J. Long, T. Darrell, Fully convolutional networks for semantic segmentation, IEEE Transactions on Pattern Analysis and Machine Intelligence 39 (4) (2017) 640–651.

[53] A. Jungo, R. Meier, E. Ermis, E. Herrmann, M. Reyes, Uncertainty-driven sanity check: Application to postoperative brain tumor cavity segmentation, arXiv preprint arXiv:1806.03106 (2018).

[54] T. Vandal, M. Livingston, C. Piho, S. Zimmerman, Prediction and uncertainty quantification of daily airport flight delays, in: International Conference on Predictive Applications and APIs, 2018, pp. 45–51.

[55] T. DeVries, G. W. Taylor, Leveraging uncertainty estimates for predicting segmentation quality, ArXiv abs/1807.00502 (2018).

[56] A. Tousignant, P. Lemaître, D. Precup, D. L. Arnold, T. Arbel, Prediction of disease progression in multiple sclerosis patients using deep learning analysis of mri data, in: International Conference on Medical Imaging with Deep Learning, 2019, pp. 483–492.

[57] A. Norouzi, A. Emami, K. Najarian, N. Karimi, S. samavi, S. M. R. Soroushmehr, Exploiting uncertainty of deep neural networks for improving segmentation accuracy in mri images, in: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2019, pp. 2322–2326.

[58] A. G. Roy, S. Conjeti, N. Navab, C. Wachinger, A. D. N. Initiative, et al., Bayesian quicknat: model uncertainty in deep whole-brain segmentation for structure-wise quality control, NeuroImage 195 (2019) 11–22.

[59] A. Filos, S. Farquhar, A. N. Gomez, T. G. J. Rudner, Z. Kenton, L. Smith, M. Alizadeh, A. de Kroon, Y. Gal, A systematic comparison of bayesian deep learning robustness in diabetic retinopathy tasks (2019). arXiv:1912.10481.

[60] R. Harper, J. Southern, A bayesian deep learning framework for end-to-end prediction of emotion from heartbeat, IEEE Transactions on Affective Computing (2020).

[61] H. P. Do, Y. Guo, A. J. Yoon, K. S. Nayak, Accuracy, uncertainty, and adaptability of automatic myocardial asl segmentation using deep cnn, Magnetic Resonance in Medicine 83 (5) (2020) 1863–1874.

[62] O. Ronneberger, P. Fischer, T. Brox, U-net: Convolutional networks for biomedical image segmentation (2015). arXiv:1505.04597.

[63] M. Teye, H. Azizpour, K. Smith, Bayesian uncertainty estimation for batch normalized deep networks, in: J. Dy, A. Krause (Eds.), Proceedings of the 35th International Conference on Machine Learning, Vol. 80 of Proceedings of Machine Learning Research, PMLR, Stockholmsmässan, Stockholm Sweden, 2018, pp. 4907–4916.

[64] C. Li, C. Chen, Y. Pu, R. Henao, L. Carin, Communication-efficient stochastic gradient mcmc for neural networks, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 33, 2019, pp. 4173–4180.

[65] A. Y. Foong, D. R. Burt, Y. Li, R. E. Turner, On the expressiveness of approximate inference in bayesian neural networks, arXiv (2019) arXiv–1909.

[66] M. Ng, F. Guo, L. Biswas, G. A. Wright, Estimating uncertainty in neural networks for segmentation quality control, Tech. rep., Technical Report. URL: https://www. doc. ic. ac. uk/˜ bglocker/public ... (2018).

[67] A. Siddhant, Z. C. Lipton, Deep Bayesian active learning for natural language processing: Results of a large-scale empirical study, in: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Brussels, Belgium, 2018, pp. 2904–2909.

[68] J. Mukhoti, P. Stenetorp, Y. Gal, On the importance of strong baselines in bayesian deep learning, arXiv preprint arXiv:1811.09385 (2018).

[69] N. Kennamer, A. T. Ihler, D. Kirkby, Empirical study of mc-dropout in various astronomical observing conditions, in: CVPR Workshops, 2019.

[70] M. A. Kupinski, J. W. Hoppin, E. Clarkson, H. H. Barrett, Ideal-observer computation in medical imaging with use of markov-chain monte carlo techniques, JOSA A 20 (3) (2003) 430–438.

[71] R. Salakhutdinov, A. Mnih, Bayesian probabilistic matrix factorization using markov chain monte carlo, in: Proceedings of the International Conference on Machine Learning, ICML '08, Association for Computing Machinery, 2008, p. 880–887.

[72] T. Salimans, D. Kingma, M. Welling, Markov chain monte carlo and variational inference: Bridging the gap, in: International Conference on Machine Learning, 2015, pp. 1218–1226.

[73] T. Chen, E. Fox, C. Guestrin, Stochastic gradient hamiltonian monte carlo, in: International conference on machine learning, 2014, pp. 1683–1691.

[74] N. Ding, Y. Fang, R. Babbush, C. Chen, R. D. Skeel, H. Neven, Bayesian sampling using stochastic gradient thermostats, in: Advances in neural information processing systems, 2014, pp. 3203–3211.

[75] C. Chen, N. Ding, L. Carin, On the convergence of stochastic gradient mcmc algorithms with high-order integrators, in: Advances in Neural Information Processing Systems, 2015, pp. 2278–2286.

[76] C. Li, A. Stevens, C. Chen, Y. Pu, Z. Gan, L. Carin, Learning weight uncertainty with stochastic gradient mcmc for shape classification, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 5666–5675.

[77] W. Gong, S. Tschiatschek, S. Nowozin, R. E. Turner, J. M. Hernández-Lobato, C. Zhang, Icebreaker: Element-wise efficient information acquisition with a bayesian deep latent gaussian model, in: Advances in Neural Information Processing Systems 32, 2019, pp. 14820–14831.

[78] Y. W. Teh, A. H. Thiery, S. J. Vollmer, Consistency and fluctuations for stochastic gradient langevin dynamics, The Journal of Machine Learning Research 17 (1) (2016) 193–225.

[79] A. Choromanska, M. Henaff, M. Mathieu, G. B. Arous, Y. LeCun, The loss surfaces of multilayer networks, in: Artificial intelligence and statistics, 2015, pp. 192–204.

[80] R. Zhang, C. Li, J. Zhang, C. Chen, A. G. Wilson, Cyclical stochastic gradient mcmc for bayesian deep learning, arXiv preprint arXiv:1902.03932 (2019).

[81] R. Luo, J. Wang, Y. Yang, J. WANG, Z. Zhu, Thermostat-assisted continuously-tempered hamiltonian monte carlo for bayesian learning, in: S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, R. Garnett (Eds.), Advances in Neural Information Processing Systems 31, Curran Associates, Inc., 2018, pp. 10673–10682.

[82] S. Duane, A. D. Kennedy, B. J. Pendleton, D. Roweth, Hybrid monte carlo, Physics letters B 195 (2) (1987) 216–222.

[83] S. Hernández, D. Vergara, M. Valdenegro-Toro, F. Jorquera, Improving predictive uncertainty estimation using dropout–hamiltonian monte carlo, Soft Computing 24 (6) (2020) 4307–4322.

[84] W. G. Hoover, Canonical dynamics: Equilibrium phase-space distributions, Physical review A 31 (3) (1985) 1695.

[85] S. Nosé, A unified formulation of the constant temperature molecular dynamics methods, The Journal of chemical physics 81 (1) (1984) 511–519.

[86] M. Welling, Y. W. Teh, Bayesian learning via stochastic gradient langevin dynamics, in: Proceedings of the 28th international conference on machine learning (ICML-11), 2011, pp. 681–688.

[87] W. Zhou, M. A. Anastasio, Markov-chain monte carlo approximation of the ideal observer using generative adversarial networks, in: Medical Imaging 2020: Image Perception, Observer Performance, and Technology Assessment, Vol. 11316, International Society for Optics and Photonics, 2020, p. 113160D.

[88] J. Kwon, Robust visual tracking based on variational auto-encoding markov chain monte carlo, Information Sciences 512 (2020) 1308 – 1323.

[89] J. Swiatkowski, K. Roth, B. S. Veeling, L. Tran, J. V. Dillon, S. Mandt, J. Snoek, T. Salimans, R. Jenatton, S. Nowozin, The k-tied normal distribution: A compact parameterization of gaussian mean field posteriors in bayesian neural networks, arXiv preprint arXiv:2002.02655 (2020).

[90] K. Posch, J. Steinbrener, J. Pilz, Variational inference to measure model uncertainty in deep neural networks, arXiv preprint arXiv:1902.10189 (2019).

[91] K. Posch, J. Pilz, Correlated parameters to accurately measure uncertainty in deep neural networks (2019). arXiv:1904.01334.

[92] H. Robbins, An empirical bayes approach to statistics, in: Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics, University of California Press, 1956, pp. 157–163.

[93] R. Krishnan, M. Subedar, O. Tickoo, Efficient priors for scalable variational inference in bayesian deep neural networks, in: 2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW), 2019, pp. 773–777.

[94] R. Krishnan, M. Subedar, O. Tickoo, A. Filos, Y. Gal, Improving mfvi in bayesian neural networks with empirical bayes: a study with diabetic retinopathy diagnosis, in: Fourht workshop on Bayesian Deep Learning (NeurIPS 2019), 2019.

[95] M. Subedar, R. Krishnan, P. L. Meyer, O. Tickoo, J. Huang, Uncertainty-aware audiovisual activity recognition using deep bayesian variational inference, in: 2019 IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 6300–6309.

[96] D. L. Marino, M. Manic, Modeling and planning under uncertainty using deep neural networks, IEEE Transactions on Industrial Informatics 15 (8) (2019) 4442–4454.

[97] T. D. Le, R. Noumeir, H. L. Quach, J. H. Kim, J. H. Kim, H. M. Kim, Critical temperature prediction for a superconductor: A variational bayesian neural network approach, IEEE Transactions on Applied Superconductivity 30 (4) (2020) 1–5.

[98] C. Louizos, M. Welling, Multiplicative normalizing flows for variational bayesian neural networks, in: Proceedings of the 34th International Conference on Machine Learning-Volume 70, JMLR. org, 2017, pp. 2218–2227.

[99] D. P. Kingma, M. Welling, Auto-encoding variational bayes (2013). arXiv:1312.6114.

[100] A. Hubin, G. Storvik, Combining model and parameter uncertainty in bayesian neural networks, arXiv preprint arXiv:1903.07594 (2019).

[101] Y. Liu, H. Qin, Z. Zhang, S. Pei, Z. Jiang, Z. Feng, J. Zhou, Probabilistic spatiotemporal wind speed forecasting based on a variational bayesian deep learning model, Applied Energy 260 (2020) 114259.

[102] S. Ryu, Y. Kwon, W. Y. Kim, A bayesian graph convolutional network for reliable prediction of molecular properties with uncertainty quantification, Chemical Science 10 (36) (2019) 8438–8446.

[103] S. Farquhar, L. Smith, Y. Gal, Try depth instead of weight correlations: Mean-field is a less restrictive assumption for variational inference in deep networks, in: Bayesian Deep Learning Workshop at NeurIPS, 2020.

[104] H. S. Hossain, M. A. A. H. Khan, N. Roy, Active learning enabled activity recognition, Pervasive and Mobile Computing 38 (2017) 312–330.

[105] S. Tong, Active learning: theory and applications, Vol. 1, Stanford University USA, 2001.

[106] Y. Gal, R. Islam, Z. Ghahramani, Deep bayesian active learning with image data, in: Proceedings of the 34th International Conference on Machine Learning-Volume 70, JMLR. org, 2017, pp. 1183–1192.

[107] N. Houlsby, F. Huszár, Z. Ghahramani, M. Lengyel, Bayesian active learning for classification and preference learning (2011). arXiv:1112.5745.

[108] R. Krishnan, M. Subedar, O. Tickoo, Specifying weight priors in bayesian deep neural networks with empirical bayes, arXiv:1906.05323 (2019). arXiv:arXiv:1906.05323.

[109] A. Kirsch, J. van Amersfoort, Y. Gal, Batchbald: Efficient and diverse batch acquisition for deep bayesian active learning (2019). arXiv:1906.08158.

[110] S. Burkhardt, J. Siekiera, S. Kramer, Semisupervised bayesian active learning for text classification, in: Bayesian Deep Learning Workshop at NeurIPS, 2018.

[111] G. K. Gudur, P. Sundaramoorthy, V. Umaashankar, Activeharnet: Towards on-device deep bayesian active learning for human activity recognition, in: The 3rd International Workshop on Deep Learning for Mobile Systems and Applications, 2019, pp. 7–12.

[112] R. Matthias, K. Karsten, G. Hanno, Deep bayesian active semi-supervised learning, in: IEEE International Conference on Machine Learning and Applications (ICMLA), IEEE, 2018, pp. 158–164.

[113] A. P. Dempster, N. M. Laird, D. B. Rubin, Maximum likelihood from incomplete data via the em algorithm, Journal of the Royal Statistical Society: Series B (Methodological) 39 (1) (1977) 1–22.

[114] M. L. di Scandalea, C. S. Perone, M. Boudreau, J. Cohen-Adad, Deep active learning for axon-myelin segmentation on histology data (2019). arXiv:1907.05143.

[115] J. Zeng, A. Lesnikowski, J. M. Alvarez, The relevance of bayesian layer positioning to model uncertainty in deep bayesian active learning, arXiv preprint arXiv:1811.12535 (2018).

[116] J. Huggins, T. Campbell, T. Broderick, Coresets for scalable bayesian logistic regression, in: Advances in Neural Information Processing Systems, 2016, pp. 4080–4088.

[117] R. Pinsler, J. Gordon, E. Nalisnick, J. M. Hernández-Lobato, Bayesian batch active learning as sparse subset approximation, in: Advances in Neural Information Processing Systems, 2019, pp. 6356–6367.

[118] M. Servajean, A. Joly, D. Shasha, J. Champ, E. Pacitti, Crowdsourcing thousands of specialized labels: A bayesian active training approach, IEEE Transactions on Multimedia 19 (6) (2017) 1376–1391.

[119] E. Simpson, S. Roberts, Bayesian methods for intelligent task assignment in crowdsourcing systems, in: Decision Making: Uncertainty, Imperfection, Deliberation and Scalability, Springer, 2015, pp. 1–32.

[120] J. Gordon, J. M. Hernández-Lobato, Bayesian semisupervised learning with deep generative models, arXiv preprint arXiv:1706.09751 (2017).

[121] T. Tran, T.-T. Do, I. Reid, G. Carneiro, Bayesian generative active deep learning, arXiv preprint arXiv:1904.11643 (2019).

[122] A. Akbari, R. Jafari, Personalizing activity recognition models with quantifying different types of uncertainty using wearable sensors, IEEE Transactions on Biomedical Engineering (2020).

[123] C. Blundell, J. Cornebise, K. Kavukcuoglu, D. Wierstra, Weight uncertainty in neural networks, arXiv preprint arXiv:1505.05424 (2015).

[124] M. Fortunato, C. Blundell, O. Vinyals, Bayesian recurrent neural networks, arXiv preprint arXiv:1704.02798 (2017).

[125] S. Ebrahimi, M. Elhoseiny, T. Darrell, M. Rohrbach, Uncertainty-guided continual learning with bayesian neural networks, arXiv preprint arXiv:1906.02425 (2019).

[126] M. de la Riva, P. Mettes, Bayesian 3d convnets for action recognition from few examples, in: Proceedings of the IEEE International Conference on Computer Vision Workshops, 2019, pp. 0–0.

[127] S. Z. Dadaneh, S. Boluki, M. Yin, M. Zhou, X. Qian, Pairwise supervised hashing with bernoulli variational auto-encoder and self-control gradient estimator, arXiv preprint arXiv:2005.10477 (2020).

[128] P. Ghosh, M. S. M. Sajjadi, A. Vergari, M. Black, B. Schölkopf, From variational to deterministic autoencoders (2019). arXiv:1903.12436.

[129] V. Edupuganti, M. Mardani, J. Cheng, S. Vasanawala, J. Pauly, Uncertainty analysis of vae-gans for compressive medical imaging, arXiv preprint arXiv:1901.11228 (2019).

[130] V. Böhm, F. Lanusse, U. Seljak, Uncertainty quantification with generative models, arXiv preprint arXiv:1910.10046 (2019).

[131] U. Seljak, B. Yu, Posterior inference unchained with EL_2O (2019). arXiv:1901.04454.

[132] L. Jin, H. Lu, G. Wen, Fast uncertainty quantification of reservoir simulation with variational u-net (2019). arXiv:1907.00718.

[133] P. Esser, E. Sutter, B. Ommer, A variational u-net for conditional appearance and shape generation (2018). arXiv:1804.04694.

[134] K. Yi, Y. Guo, Y. Fan, J. Hamann, Y. G. Wang, Cosmovae: Variational autoencoder for cmb image inpainting (2020). arXiv:2001.11651.

[135] N. Mehrasa, A. A. Jyothi, T. Durand, J. He, L. Sigal, G. Mori, A variational auto-encoder model for stochastic point processes (2019). arXiv:1904.03273.

[136] K. Sato, K. Hama, T. Matsubara, K. Uehara, Predictable uncertainty-aware unsupervised deep anomaly segmentation, in: 2019 International Joint Conference on Neural Networks (IJCNN), 2019, pp. 1–7.

[137] S. Mishra, S. Flaxman, S. Bhatt, πvae: Encoding stochastic process priors with variational autoencoders (2020). arXiv:2002.06873.

[138] M. Garnelo, D. Rosenbaum, C. J. Maddison, T. Ramalho, D. Saxton, M. Shanahan, Y. W. Teh, D. J. Rezende, S. Eslami, Conditional neural processes, arXiv preprint arXiv:1807.01613 (2018).

[139] F. Guo, R. Xie, B. Huang, A deep learning just-in-time modeling approach for soft sensor based on variational autoencoder, Chemometrics and Intelligent Laboratory Systems 197 (2020) 103922.

[140] E. Daxberger, J. M. Hernández-Lobato, Bayesian variational autoencoders for unsupervised out-of-distribution detection (2019). arXiv:1912.05651.

[141] D. J. MacKay, D. J. Mac Kay, Information theory, inference and learning algorithms, Cambridge university press, 2003.

[142] H. Ritter, A. Botev, D. Barber, A scalable laplace approximation for neural networks, in: 6th International Conference on Learning Representations, ICLR 2018-Conference Track Proceedings, Vol. 6, International Conference on Representation Learning, 2018.

[143] J. Feng, M. Durner, Z.-C. Marton, F. Balint-Benczedi, R. Triebel, Introspective robot perception using smoothed predictions from bayesian neural networks (2019).

[144] K. Shinde, J. Lee, M. Humt, A. Sezgin, R. Triebel, Learning multiplicative interactions with bayesian neural networks for visual-inertial odometry, arXiv preprint arXiv:2007.07630 (2020).

[145] J. Lee, M. Humt, J. Feng, R. Triebel, Estimating model uncertainty of neural networks in sparse information form, arXiv preprint arXiv:2006.11631 (2020).

[146] M. Humt, J. Lee, R. Triebel, Bayesian optimization meets laplace approximation for robotic introspection, arXiv preprint arXiv:2010.16141 (2020).

[147] T. Doan, B. Mazoure, A. Durand, J. Pineau, R. D. Hjelm, Attraction-repulsion actor-critic for continuous control reinforcement learning, arXiv preprint arXiv:1909.07543 (2019).

[148] X. Zhao, S. Hu, J.-H. Cho, F. Chen, Uncertainty-based decision making using deep reinforcement learning, in: 2019 22th International Conference on Information Fusion (FUSION), IEEE, 2019, pp. 1–8.

[149] G. Lee, B. Hou, A. Mandalika, J. Lee, S. Choudhury, S. S. Srinivasa, Bayesian policy optimization for model uncertainty, arXiv preprint arXiv:1810.01014 (2018).

[150] B. O'Donoghue, I. Osband, R. Munos, V. Mnih, The uncertainty bellman equation and exploration, arXiv preprint arXiv:1709.05380 (2017).

[151] G. Kahn, A. Villaflor, V. Pong, P. Abbeel, S. Levine, Uncertainty-aware reinforcement learning for collision avoidance, arXiv preprint arXiv:1702.01182 (2017).

[152] M. Ghavamzadeh, S. Mannor, J. Pineau, A. Tamar, et al., Bayesian reinforcement learning: A survey, Foundations and Trends® in Machine Learning 8 (5-6) (2015) 359–483.

[153] A. Tschantz, B. Millidge, A. K. Seth, C. L. Buckley, Reinforcement learning through active inference, arXiv preprint arXiv:2002.12636 (2020).

[154] G. Kalweit, J. Boedecker, Uncertainty-driven imagination for continuous deep reinforcement learning, in: Conference on Robot Learning, 2017, pp. 195–206.

[155] C. Tegho, P. Budzianowski, M. Gašić, Benchmarking uncertainty estimates with deep reinforcement learning for dialogue policy optimisation, in: 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2018, pp. 6069–6073.

[156] C. Riquelme, G. Tucker, J. Snoek, Deep bayesian bandits showdown: An empirical comparison of bayesian deep networks for thompson sampling, arXiv preprint arXiv:1802.09127 (2018).

[157] D. Janz, J. Hron, P. Mazur, K. Hofmann, J. M. Hernández-Lobato, S. Tschiatschek, Successor uncertainties: exploration and uncertainty in temporal difference learning, in: Advances in Neural Information Processing Systems, 2019, pp. 4509–4518.

[158] M. Shen, J. P. How, Active perception in adversarial scenarios using maximum entropy deep reinforcement learning, in: 2019 International Conference on Robotics and Automation (ICRA), IEEE, 2019, pp. 3384–3390.

[159] M. Benatan, E. O. Pyzer-Knapp, Fully bayesian recurrent neural networks for safe reinforcement learning, arXiv preprint arXiv:1911.03308 (2019).

[160] W. Huang, J. Zhang, K. Huang, Bootstrap estimated uncertainty of the environment model for model-based reinforcement learning, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 33, 2019, pp. 3870–3877.

[161] H. Eriksson, C. Dimitrakakis, Epistemic risk-sensitive reinforcement learning, arXiv preprint arXiv:1906.06273 (2019).

[162] B. Lötjens, M. Everett, J. P. How, Safe reinforcement learning with model uncertainty estimates, in: 2019 International Conference on Robotics and Automation (ICRA), IEEE, 2019, pp. 8662–8668.

[163] W. R. Clements, B.-M. Robaglia, B. Van Delft, R. B. Slaoui, S. Toth, Estimating risk and uncertainty in deep reinforcement learning, arXiv preprint arXiv:1905.09638 (2019).

[164] A. M. Metelli, A. Likmeta, M. Restelli, Propagating uncertainty in reinforcement learning via wasserstein barycenters, in: Advances in Neural Information Processing Systems, 2019, pp. 4333–4345.

[165] C. D'Eramo, A. Cini, M. Restelli, Exploiting action-value uncertainty to drive exploration in reinforcement learning, in: 2019 International Joint Conference on Neural Networks (IJCNN), IEEE, 2019, pp. 1–8.

[166] A. Tschantz, M. Baltieri, A. K. Seth, C. L. Buckley, Scaling active inference, in: 2020 International Joint Conference on Neural Networks (IJCNN), IEEE, 2020, pp. 1–8.

[167] H.-S. Lee, Y. Zhang, W. Zame, C. Shen, J.-W. Lee, M. van der Schaar, Robust recursive partitioning for heterogeneous treatment effects with uncertainty quantification, Advances in Neural Information Processing Systems 33 (2020).

[168] R. Ramakrishnan, E. Kamar, D. Dey, E. Horvitz, J. Shah, Blind spot detection for safe sim-to-real transfer, Journal of Artificial Intelligence Research 67 (2020) 191–234.

[169] J. Zhang, B. Kailkhura, T. Y.-J. Han, Mix-n-match: Ensemble and compositional methods for uncertainty calibration in deep learning, in: International Conference on Machine Learning, PMLR, 2020, pp. 11117–11128.

[170] J. Z. Liu, Variable selection with rigorous uncertainty quantification using bayesian deep neural networks, in: Bayesian Deep Learning Workshop at NeurIPS, 2019.

[171] B. Lakshminarayanan, A. Pritzel, C. Blundell, Simple and scalable predictive uncertainty estimation using deep ensembles, in: Advances in neural information processing systems, 2017, pp. 6402–6413.

[172] S. Jain, G. Liu, J. Mueller, D. Gifford, Maximizing overall diversity for improved uncertainty estimates in deep ensembles, arXiv preprint arXiv:1906.07380 (2019).

[173] F. K. Gustafsson, M. Danelljan, T. B. Schön, Evaluating scalable bayesian deep learning methods for robust computer vision, arXiv preprint arXiv:1906.01620 (2019).

[174] K. Chua, R. Calandra, R. McAllister, S. Levine, Deep reinforcement learning in a handful of trials using probabilistic dynamics models, in: Advances in Neural Information Processing Systems, 2018, pp. 4754–4765.

[175] J. Liu, J. Paisley, M.-A. Kioumourtzoglou, B. Coull, Accurate uncertainty estimation and decomposition in ensemble learning, in: Advances in Neural Information Processing Systems, 2019, pp. 8950–8961.

[176] A. Malinin, M. Gales, Uncertainty in structured prediction, arXiv preprint arXiv:2002.07650 (2020).

[177] A. Malinin, B. Mlodozeniec, M. Gales, Ensemble distribution distillation, arXiv preprint arXiv:1905.00076 (2019).

[178] Y. Wen, D. Tran, J. Ba, Batchensemble: an alternative approach to efficient ensemble and lifelong learning, arXiv preprint arXiv:2002.06715 (2020).

[179] R. Hu, Q. Huang, S. Chang, H. Wang, J. He, The mbpep: a deep ensemble pruning algorithm providing high quality uncertainty prediction, Applied Intelligence 49 (8) (2019) 2942–2955.

[180] P. L. McDermott, C. K. Wikle, Deep echo state networks with uncertainty quantification for spatio-temporal forecasting, Environmetrics 30 (3) (2019) e2553.

[181] A. Ashukha, A. Lyzhov, D. Molchanov, D. Vetrov, Pitfalls of in-domain uncertainty estimation and ensembling in deep learning, arXiv preprint arXiv:2002.06470 (2020).

[182] G. Wang, W. Li, S. Ourselin, T. Vercauteren, Automatic brain tumor segmentation using convolutional neural networks with test-time augmentation, in: International MICCAI Brainlesion Workshop, Springer, 2018, pp. 61–72.

[183] A. Lyzhov, Y. Molchanova, A. Ashukha, D. Molchanov, D. Vetrov, Greedy policy search: A simple baseline for learnable test-time augmentation, in: Conference on Uncertainty in Artificial Intelligence, PMLR, 2020, pp. 1308–1317.

[184] D. Shanmugam, D. Blalock, G. Balakrishnan, J. Guttag, When and why test-time augmentation works, arXiv preprint arXiv:2011.11156 (2020).

[185] A. G. Wilson, P. Izmailov, Bayesian deep learning and a probabilistic perspective of generalization, in: Advances in Neural Information Processing Systems 31, 2020, pp. 1–12.

[186] R. Pop, P. Fulop, Deep ensemble bayesian active learning: Addressing the mode collapse issue in monte carlo dropout via ensembles, arXiv preprint arXiv:1811.03897 (2018).

[187] E. A. Antonelo, E. Camponogara, B. Foss, Echo state networks for data-driven downhole pressure estimation in gas-lift oil wells, Neural Networks 85 (2017) 106–117.

[188] Q. Ma, L. Shen, G. W. Cottrell, Deep-esn: A multiple projection-encoding hierarchical reservoir computing framework, arXiv preprint arXiv:1711.05255 (2017).

[189] E. Fersini, E. Messina, F. A. Pozzi, Sentiment analysis: Bayesian ensemble learning, Decision support systems 68 (2014) 26–38.

[190] T. Pearce, M. Zaki, A. Brintrup, N. Anastassacos, A. Neely, Uncertainty in neural networks: Bayesian ensembling, arXiv preprint arXiv:1810.05546 (2018).

[191] R. Pop, P. Fulop, Deep ensemble bayesian active learning, in: Bayesian Deep Learning Workshop at NeurIPS, 2020.

[192] T. Pearce, F. Leibfried, A. Brintrup, M. Zaki, A. Neely, Uncertainty in neural networks: Approximately bayesian ensembling, in: The 23rd International Conference on Artificial Intelligence and Statistics, AISTATS 2020, 2020.

[193] T. Tsiligkaridis, Information aware max-norm dirichlet networks for predictive uncertainty estimation, Neural Networks 135 (2021) 105–114.

[194] M. Sensoy, L. Kaplan, M. Kandemir, Evidential deep learning to quantify classification uncertainty, in: Advances in Neural Information Processing Systems, 2018, pp. 3179–3189.

[195] R. Ardywibowo, S. Boluki, X. Gong, Z. Wang, X. Qian, Nads: Neural architecture distribution search for uncertainty awareness, arXiv preprint arXiv:2006.06646 (2020).

[196] H. Kabir, A. Khosravi, A. Kavousi-Fard, S. Nahavandi, D. Srinivasan, Optimal uncertainty-guided neural network training, arXiv preprint arXiv:1912.12761 (2019).

[197] Y. Geifman, G. Uziel, R. El-Yaniv, Bias-reduced uncertainty estimation for deep neural classifiers, arXiv preprint arXiv:1805.08206 (2018).

[198] N. Tagasovska, D. Lopez-Paz, Single-model uncertainties for deep learning, in: Advances in Neural Information Processing Systems, 2019, pp. 6414–6425.

[199] J. van Amersfoort, L. Smith, Y. W. Teh, Y. Gal, Simple and scalable epistemic uncertainty estimation using a single deep deterministic neural network, arXiv preprint arXiv:2003.02037 (2020).

[200] J. van Amersfoort, L. Smith, Y. W. Teh, Y. Gal, Uncertainty estimation using a single deep deterministic neural network, in: Proceedings of the 37th International Conference on Machine Learning, 2020.

[201] N. Tagasovska, D. Lopez-Paz, Frequentist uncertainty estimates for deep learning, arXiv preprint arXiv:1811.00908 (2018).

[202] A. Mobiny, H. V. Nguyen, S. Moulik, N. Garg, C. C. Wu, Dropconnect is effective in modeling uncertainty of bayesian deep networks, arXiv preprint arXiv:1906.04569 (2019).

[203] Y. Chen, Y. Xue, Y. Ma, F. Xu, Measures of uncertainty for neighborhood rough sets, Knowledge-Based Systems 120 (2017) 226–235.

[204] V. Kuleshov, N. Fenner, S. Ermon, Accurate uncertainties for deep learning using calibrated regression, arXiv preprint arXiv:1807.00263 (2018).

[205] T. G. Rudner, F. Wenzel, Y. W. Teh, Y. Gal, The natural neural tangent kernel: Neural network training dynamics under natural gradient descent, in: Fourht workshop on Bayesian Deep Learning (NeurIPS 2019), 2019.

[206] B. N. Patro, M. Lunayach, S. Patel, V. P. Namboodiri, U-cam: Visual explanation using uncertainty based class activation maps, in: Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 7444–7453.

[207] S. Depeweg, J. M. Hernández-Lobato, F. Doshi-Velez, S. Udluft, Decomposition of uncertainty in bayesian deep learning for efficient and risk-sensitive learning, arXiv preprint arXiv:1710.07283 (2017).

[208] R. Y. Rohekar, Y. Gurwicz, S. Nisimov, G. Novik, Modeling uncertainty by learning a hierarchy of deep neural connections, in: Advances in Neural Information Processing Systems, 2019, pp. 4246–4256.

[209] S. Khan, M. Hayat, S. W. Zamir, J. Shen, L. Shao, Striking the right balance with uncertainty, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 103–112.

[210] N. Pawlowski, A. Brock, M. C. Lee, M. Rajchl, B. Glocker, Implicit weight uncertainty in neural networks, arXiv preprint arXiv:1711.01297 (2017).

[211] X. Wang, Y. Luo, D. Crankshaw, A. Tumanov, F. Yu, J. E. Gonzalez, Idk cascades: Fast deep learning by learning not to overthink, arXiv preprint arXiv:1706.00885 (2017).

[212] Y. Yang, P. Perdikaris, Adversarial uncertainty quantification in physics-informed neural networks, Journal of Computational Physics 394 (2019) 136–152.

[213] F. Brickwedde, S. Abraham, R. Mester, Mono-sf: Multi-view geometry meets single-view depth for monocular scene flow estimation of dynamic traffic scenes, in: Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 2780–2790.

[214] S. Thulasidasan, G. Chennupati, J. A. Bilmes, T. Bhattacharya, S. Michalak, On mixup training: Improved calibration and predictive uncertainty for deep neural networks, in: Advances in Neural Information Processing Systems, 2019, pp. 13888–13899.

[215] Y. Zhang, K. Song, Y. Sun, S. Tan, M. Udell, why should you trust my explanation?" understanding uncertainty in lime explanations, arXiv preprint arXiv:1904.12991 (2019).

[216] F. Sheikholeslami, S. Jain, G. B. Giannakis, Minimum uncertainty based detection of adversaries in deep neural networks, arXiv preprint arXiv:1904.02841 (2019).

[217] Y. Li, L. Liu, R. T. Tan, Decoupled certainty-driven consistency loss for semi-supervised learning, arXiv (2019) arXiv–1901.

[218] E. Englesson, H. Azizpour, Efficient evaluation-time uncertainty estimation by improved distillation, arXiv preprint arXiv:1906.05419 (2019).

[219] B. C. Csáji, K. B. Kis, Distribution-free uncertainty quantification for kernel methods by gradient perturbations, Machine Learning 108 (8-9) (2019) 1677–1699.

[220] D. Hendrycks, M. Mazeika, S. Kadavath, D. Song, Using self-supervised learning can improve model robustness and uncertainty, in: Advances in Neural Information Processing Systems, 2019, pp. 15663–15674.

[221] B. S. Veeling, R. v. d. Berg, M. Welling, Predictive uncertainty through quantization, arXiv preprint arXiv:1810.05500 (2018).

[222] F. J. Bragman, R. Tanno, Z. Eaton-Rosen, W. Li, D. J. Hawkes, S. Ourselin, D. C. Alexander, J. R. McClelland, M. J. Cardoso, Quality control in radiotherapy-treatment planning using multi-task learning and uncertainty estimation, in: International Conference on Medical Imaging with Deep Learning, 2018.

[223] S. Stanton, W. Maddox, I. Delbridge, A. G. Wilson, Kernel interpolation for scalable online gaussian processes, in: International Conference on Artificial Intelligence and Statistics, PMLR, 2021, pp. 3133–3141.

[224] A. Mishkin, F. Kunstner, D. Nielsen, M. Schmidt, M. E. Khan, Slang: Fast structured covariance approximations for bayesian deep learning with natural gradient, in: Advances in Neural Information Processing Systems, 2018, pp. 6245–6255.

[225] A. Malinin, M. Gales, Predictive uncertainty estimation via prior networks, in: Advances in Neural Information Processing Systems, 2018, pp. 7047–7058.

[226] H. Yang, J. Zhang, H. Dong, N. Inkawhich, A. Gardner, A. Touchet, W. Wilkes, H. Berry, H. Li, Dverge: diversifying vulnerabilities for enhanced robust generation of ensembles, arXiv preprint arXiv:2009.14720 (2020).

[227] M. Jain, S. Lahlou, H. Nekoei, V. Butoi, P. Bertin, J. Rector-Brooks, M. Korablyov, Y. Bengio, Deup: Direct epistemic uncertainty prediction, arXiv preprint arXiv:2102.08501 (2021).

[228] E. Daxberger, E. Nalisnick, J. U. Allingham, J. Antorán, J. M. Hernández-Lobato, Expressive yet tractable bayesian deep learning via subnetwork inference, arXiv preprint arXiv:2010.14689 (2020).

[229] X. Zhao, F. Chen, S. Hu, et al., Uncertainty-aware prediction for graph neural networks (2019).

[230] Q. Wang, H. Van Hoof, Doubly stochastic variational inference for neural processes with hierarchical latent variables, in: International Conference on Machine Learning, PMLR, 2020, pp. 10018–10028.

[231] A. Amini, W. Schwarting, A. Soleimany, D. Rus, Deep evidential regression, arXiv preprint arXiv:1910.02600 (2019).

[232] T. Tsiligkaridis, Failure prediction by confidence estimation of uncertainty-aware dirichlet networks, arXiv preprint arXiv:2010.09865 (2020).

[233] C. Corbière, N. Thome, A. Bar-Hen, M. Cord, P. Pérez, Addressing failure prediction by learning model confidence, in: Advances in Neural Information Processing Systems, 2019, pp. 2902–2913.

[234] A. Damianou, N. Lawrence, Deep gaussian processes, in: Artificial Intelligence and Statistics, 2013, pp. 207–215.

[235] D. Duvenaud, O. Rippel, R. Adams, Z. Ghahramani, Avoiding pathologies in very deep networks, in: Artificial Intelligence and Statistics, 2014, pp. 202–210.

[236] H. Salimbeni, M. Deisenroth, Doubly stochastic variational inference for deep gaussian processes, in: Advances in Neural Information Processing Systems, 2017, pp. 4588–4599.

[237] A. Borovykh, A gaussian process perspective on convolutional neural networks, arXiv preprint arXiv:1810.10798 (2018).

[238] H. Yu, Y. Chen, Z. Dai, K. H. Low, P. Jaillet, Implicit posterior variational inference for deep gaussian processes, arXiv preprint arXiv:1910.11998 (2019).

[239] S. Sun, W. Dong, Q. Liu, Multi-view representation learning with deep gaussian processes, IEEE Transactions on Pattern Analysis and Machine Intelligence (2020).

[240] S. W. Ober, L. Aitchison, Global inducing point variational posteriors for bayesian neural networks and deep gaussian processes, arXiv preprint arXiv:2005.08140 (2020).

[241] C. E. Rasmussen, Gaussian processes for machine learning, in: Advanced Lectures on Machine Learning, MIT Press, 2006.

[242] T. Teng, J. Chen, Y. Zhang, B. K. H. Low, Scalable variational bayesian kernel selection for sparse gaussian process regression., in: AAAI, 2020, pp. 5997–6004.

[243] S. J. Oh, K. Murphy, J. Pan, J. Roth, F. Schroff, A. Gallagher, Modeling uncertainty with hedged instance embedding, arXiv preprint arXiv:1810.00319 (2018).

[244] M. Havasi, J. M. Hernández-Lobato, J. J. Murillo-Fuentes, Inference in deep gaussian processes using stochastic gradient hamiltonian monte carlo, in: Advances in neural information processing systems, 2018, pp. 7506–7516.

[245] W. Maddox, T. Garipov, P. Izmailov, D. Vetrov, A. G. Wilson, Fast uncertainty estimates and bayesian model averaging of dnns, in: Uncertainty in Deep Learning Workshop at UAI, 2018.

[246] P. Izmailov, D. Podoprikhin, T. Garipov, D. Vetrov, A. G. Wilson, Averaging weights leads to wider optima and better generalization, arXiv preprint arXiv:1803.05407 (2018).

[247] W. J. Maddox, P. Izmailov, T. Garipov, D. P. Vetrov, A. G. Wilson, A simple baseline for bayesian uncertainty in deep learning, in: Advances in Neural Information Processing Systems, 2019, pp. 13153–13164.

[248] Y. Wen, P. Vicol, J. Ba, D. Tran, R. Grosse, Flipout: Efficient pseudo-independent weight perturbations on mini-batches, arXiv preprint arXiv:1803.04386 (2018).

[249] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, R. Fergus, Intriguing properties of neural networks (2013). arXiv:1312.6199.

[250] J. Bradshaw, A. G. d. G. Matthews, Z. Ghahramani, Adversarial examples, uncertainty, and transfer testing robustness in gaussian process hybrid deep networks, arXiv preprint arXiv:1707.02476 (2017).

[251] J. Choi, D. Chun, H. Kim, H.-J. Lee, Gaussian yolov3: An accurate and fast object detector using localization uncertainty for autonomous driving, in: Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 502–511.

[252] J. Redmon, A. Farhadi, Yolov3: An incremental improvement (2018). arXiv:1804.02767.

[253] M. E. Khan, D. Nielsen, V. Tangkaratt, W. Lin, Y. Gal, A. Srivastava, Fast and scalable bayesian deep learning by weight-perturbation in adam, arXiv preprint arXiv:1806.04854 (2018).

[254] S. Sun, C. Chen, L. Carin, Learning structured weight uncertainty in bayesian neural networks, in: Artificial Intelligence and Statistics, 2017, pp. 1283–1292.

[255] A. K. Gupta, D. K. Nagar, Matrix variate distributions, Vol. 104, CRC Press, 2018.

[256] C. Louizos, M. Welling, Structured and efficient variational deep learning with matrix gaussian posteriors, in: International Conference on Machine Learning, 2016, pp. 1708–1716.

[257] M. Van der Wilk, C. E. Rasmussen, J. Hensman, Convolutional gaussian processes, in: Advances in Neural Information Processing Systems, 2017, pp. 2849–2858.

[258] K. Blomqvist, S. Kaski, M. Heinonen, Deep convolutional gaussian processes, in: Joint European Conference on Machine Learning and Knowledge Discovery in Databases, Springer, 2019, pp. 582–597.

[259] G.-L. Tran, E. V. Bonilla, J. Cunningham, P. Michiardi, M. Filippone, Calibrating deep convolutional gaussian processes, in: The 22nd International Conference on Artificial Intelligence and Statistics, 2019, pp. 1554–1563.

[260] V. Dutordoir, M. van der Wilk, A. Artemev, M. Tomczak, J. Hensman, Translation insensitivity for deep convolutional gaussian processes, arXiv preprint arXiv:1902.05888 (2019).

[261] J. Shi, M. Titsias, A. Mnih, Sparse orthogonal variational inference for gaussian processes, in: International Conference on Artificial Intelligence and Statistics, 2020, pp. 1932–1942.

[262] A. Atanov, A. Ashukha, D. Molchanov, K. Neklyudov, D. Vetrov, Uncertainty estimation via stochastic batch normalization, arXiv preprint arXiv:1802.04893 (2018).

[263] K. Neklyudov, D. Molchanov, A. Ashukha, D. Vetrov, Variance networks: When expectation does not meet your expectations, arXiv preprint arXiv:1803.03764 (2018).

[264] C. Tzelepis, V. Mezaris, I. Patras, Linear maximum margin classifier for learning from uncertain data, IEEE transactions on pattern analysis and machine intelligence 40 (12) (2017) 2948–2962.

[265] T. Kanamori, A. Takeda, T. Suzuki, Conjugate relation between loss functions and uncertainty sets in classification problems, The Journal of Machine Learning Research 14 (1) (2013) 1461–1504.

[266] T. Pereira, S. Cardoso, M. Guerreiro, S. C. Madeira, A. D. N. Initiative, et al., Targeting the uncertainty of predictions at patient-level using an ensemble of classifiers coupled with calibration methods, venn-abers, and conformal predictors: A case study in ad, Journal of Biomedical Informatics 101 (2020) 103350.

[267] I. Partalas, G. Tsoumakas, I. Vlahavas, An ensemble uncertainty aware measure for directed hill climbing ensemble pruning, Machine Learning 81 (3) (2010) 257–282.

[268] A. A. Peterson, R. Christensen, A. Khorshidi, Addressing uncertainty in atomistic machine learning, Physical Chemistry Chemical Physics 19 (18) (2017) 10978–10985.

[269] J. Guynn, Google photos labeled black people'gorillas', USA Today 1 (2015).

[270] A. Kendall, Y. Gal, What uncertainties do we need in bayesian deep learning for computer vision?, in: Advances in neural information processing systems, 2017, pp. 5574–5584.

[271] A. Taha, Y.-T. Chen, T. Misu, A. Shrivastava, L. Davis, Unsupervised data uncertainty learning in visual retrieval systems, arXiv preprint arXiv:1902.02586 (2019).

[272] G. Dorta, S. Vicente, L. Agapito, N. D. Campbell, I. Simpson, Structured uncertainty prediction networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 5477–5485.

[273] A. Asai, D. Ikami, K. Aizawa, Multi-task learning based on separable formulation of depth estimation and its uncertainty., in: CVPR Workshops, 2019, pp. 21–24.

[274] C. Liu, J. Gu, K. Kim, S. G. Narasimhan, J. Kautz, Neural rgb (r) d sensing: Depth and uncertainty from a video camera, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 10986–10995.

[275] A. Harakeh, M. Smart, S. L. Waslander, Bayesod: A bayesian approach for uncertainty estimation in deep object detectors, arXiv preprint arXiv:1903.03838 (2019).

[276] M. T. Le, F. Diehl, T. Brunner, A. Knol, Uncertainty estimation for deep neural object detectors in safety-critical applications, in: 2018 21st International Conference on Intelligent Transportation Systems (ITSC), IEEE, 2018, pp. 3873–3878.

[277] Y. He, C. Zhu, J. Wang, M. Savvides, X. Zhang, Bounding box regression with uncertainty for accurate object detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 2888–2897.

[278] P.-Y. Huang, W.-T. Hsu, C.-Y. Chiu, T.-F. Wu, M. Sun, Efficient uncertainty estimation for semantic segmentation in videos, in: Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 520–535.

[279] G. Pascual, S. Seguí, J. Vitria, Uncertainty gated network for land cover segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2018, pp. 276–279.

[280] C. Martinez, K. M. Potter, M. D. Smith, E. A. Donahue, L. Collins, J. P. Korbin, S. A. Roberts, Segmentation certainty through uncertainty: Uncertainty-refined binary volumetric segmentation under multifactor domain shift, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2019, pp. 0–0.

[281] J. Postels, F. Ferroni, H. Coskun, N. Navab, F. Tombari, Sampling-free epistemic uncertainty estimation using approximated variance propagation, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 2931–2940.

[282] E. Ilg, Ö. Çiçek, S. Galesso, A. Klein, O. Makansi, F. Hutter, T. Brox, Uncertainty estimates for optical flow with multi-hypotheses networks, arXiv preprint arXiv:1802.07095 (2018) 81.

[283] J. Gast, S. Roth, Lightweight probabilistic deep networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 3369–3378.

[284] A. Loquercio, M. Segù, D. Scaramuzza, A general framework for uncertainty estimation in deep learning, arXiv preprint arXiv:1907.06890 (2019).

[285] N. B. Gundavarapu, D. Srivastava, R. Mitra, A. Sharma, A. Jain, Structured aleatoric uncertainty in human pose estimation., in: CVPR Workshops, Vol. 2, 2019.

[286] L. Bertoni, S. Kreiss, A. Alahi, Monoloco: Monocular 3d pedestrian localization and uncertainty estimation, in: Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 6861–6871.

[287] S. Prokudin, P. Gehler, S. Nowozin, Deep directional statistics: Pose estimation with uncertainty quantification, in: Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 534–551.

[288] J. Zheng, Augmented deep representations for unconstrained still/video-based face recognition, Ph.D. thesis, University of Maryland, College Park (2019).

[289] T. Yu, D. Li, Y. Yang, T. M. Hospedales, T. Xiang, Robust person re-identification by modelling feature uncertainty, in: Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 552–561.

[290] J. Zheng, R. Yu, J.-C. Chen, B. Lu, C. D. Castillo, R. Chellappa, Uncertainty modeling of contextual-connections between tracklets for unconstrained video-based face recognition, in: Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 703–712.

[291] A. Kendall, R. Cipolla, Modelling uncertainty in deep learning for camera relocalization, in: 2016 IEEE international conference on Robotics and Automation (ICRA), IEEE, 2016, pp. 4762–4769.

[292] J. C. Peterson, R. M. Battleday, T. L. Griffiths, O. Russakovsky, Human uncertainty makes classification more robust, in: Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 9617–9626.

[293] G. Carbone, M. Wicker, L. Laurenti, A. Patane, L. Bortolussi, G. Sanguinetti, Robustness of bayesian neural networks to gradient-based attacks, arXiv preprint arXiv:2002.04359 (2020).

[294] O. Makansi, E. Ilg, O. Cicek, T. Brox, Overcoming limitations of mixture density networks: A sampling and fitting framework for multimodal future prediction, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 7144–7153.

[295] A. Bhattacharyya, M. Fritz, B. Schiele, Long-term on-board prediction of people in traffic scenes under uncertainty, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 4194–4202.

[296] A. Eldesokey, M. Felsberg, K. Holmquist, M. Persson, Uncertainty-aware cnns for depth completion: Uncertainty from beginning to end, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 12014–12023.

[297] F. K. Gustafsson, M. Danelljan, T. B. Schon, Evaluating scalable bayesian deep learning methods for robust computer vision, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 2020, pp. 318–319.

[298] M. Cai, F. Lu, Y. Sato, Generalizing hand segmentation in egocentric videos with uncertainty-guided model adaptation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 14392–14401.

[299] N. Hochgeschwender, P. Plöger, F. Kirchner, M. Valdenegro-Toro, et al., Evaluating uncertainty estimation methods on 3d semantic segmentation of point clouds, arXiv preprint arXiv:2007.01787 (2020).

[300] Y.-T. Chang, Q. Wang, W.-C. Hung, R. Piramuthu, Y.-H. Tsai, M.-H. Yang, Mixup-cam: Weakly-supervised semantic segmentation via uncertainty regularization, arXiv preprint arXiv:2008.01201 (2020).

[301] Y.-H. Huang, M. Proesmans, S. Georgoulis, L. Van Gool, Uncertainty based model selection for fast semantic segmentation, in: 2019 16th International Conference on Machine Vision Applications (MVA), IEEE, 2019, pp. 1–6.

[302] C. Sakaridis, D. Dai, L. V. Gool, Guided curriculum model adaptation and uncertainty-aware evaluation for semantic nighttime image segmentation, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 7374–7383.

[303] Z. Zheng, Y. Yang, Rectifying pseudo label learning via uncertainty estimation for domain adaptive semantic segmentation, International Journal of Computer Vision (2020) 1–15.

[304] Q. Zhou, Z. Feng, G. Cheng, X. Tan, J. Shi, L. Ma, Uncertainty-aware consistency regularization for cross-domain semantic segmentation, arXiv preprint arXiv:2004.08878 (2020).

[305] M. Ravanbakhsh, T. Klein, K. Batmanghelich, M. Nabi, Uncertainty-driven semantic segmentation through human-machine collaborative learning, arXiv preprint arXiv:1909.00626 (2019).

[306] K. Maag, M. Rottmann, H. Gottschalk, Time-dynamic estimates of the reliability of deep semantic segmentation networks, in: 2020 IEEE 32nd International Conference on Tools with Artificial Intelligence (ICTAI), IEEE, 2020, pp. 502–509.

[307] W. Ge, S. Yang, Y. Yu, Multi-evidence filtering and fusion for multi-label classification, object detection and semantic segmentation based on weakly supervised learning, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 1277–1286.

[308] H. Blum, P.-E. Sarlin, J. Nieto, R. Siegwart, C. Cadena, Fishyscapes: A benchmark for safe semantic segmentation in autonomous driving, in: Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops, 2019, pp. 0–0.

[309] M. Rottmann, M. Schubert, Uncertainty measures and prediction quality rating for the semantic segmentation of nested multi resolution street scene images, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 2019, pp. 0–0.

[310] P. Hu, S. Sclaroff, K. Saenko, Uncertainty-aware learning for zero-shot semantic segmentation, Advances in Neural Information Processing Systems 33 (2020).

[311] S. Xie, Z. Feng, Y. Chen, S. Sun, C. Ma, M. Song, Deal: Difficulty-aware active learning for semantic segmentation, in: Proceedings of the Asian Conference on Computer Vision, 2020.

[312] Y. Siddiqui, J. Valentin, M. Nießner, Viewal: Active learning with viewpoint entropy for semantic segmentation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 9433–9443.

[313] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. Van Der Laak, B. Van Ginneken, C. I. Sánchez, A survey on deep learning in medical image analysis, Medical image analysis 42 (2017) 60–88.

[314] Z. W. Lim, M. L. Lee, W. Hsu, T. Y. Wong, Building trust in deep learning system towards automated disease detection, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 33, 2019, pp. 9516–9521.

[315] M. W. Dusenberry, D. Tran, E. Choi, J. Kemp, J. Nixon, G. Jerfel, K. Heller, A. M. Dai, Analyzing the role of model uncertainty for electronic health records, in: Proceedings of the ACM Conference on Health, Inference, and Learning, 2020, pp. 204–213.

[316] Y. Xia, D. Yang, Z. Yu, F. Liu, J. Cai, L. Yu, Z. Zhu, D. Xu, A. Yuille, H. Roth, Uncertainty-aware multi-view co-training for semi-supervised medical image segmentation and domain adaptation, Medical Image Analysis (2020) 101766.

[317] R. Liu, S. Cheng, L. Tian, J. Yi, Deep spectral learning for label-free optical imaging oximetry with uncertainty quantification, Light: Science & Applications 8 (1) (2019) 1–13.

[318] S. Kohl, B. Romera-Paredes, C. Meyer, J. De Fauw, J. R. Ledsam, K. Maier-Hein, S. A. Eslami, D. J. Rezende, O. Ronneberger, A probabilistic u-net for segmentation of ambiguous images, in: Advances in Neural Information Processing Systems, 2018, pp. 6965–6975.

[319] T. Araújo, G. Aresta, L. Mendonça, S. Penas, C. Maia, Â. Carneiro, A. M. Mendonça, A. Campilho, Dr— graduate: uncertainty-aware deep learning-based diabetic retinopathy grading in eye fundus images, Medical Image Analysis (2020) 101715.

[320] D. Karimi, Q. Zeng, P. Mathur, A. Avinash, S. Mahdavi, I. Spadinger, P. Abolmaesumi, S. E. Salcudean, Accurate and robust deep learning-based segmentation of the prostate clinical target volume in ultrasound images, Medical image analysis 57 (2019) 186–196.

[321] M. Combalia, F. Hueto, S. Puig, J. Malvehy, V. Vilaplana, Uncertainty estimation in deep neural networks for dermoscopic image classification, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 2020, pp. 744–745.

[322] L. Dahal, A. Kafle, B. Khanal, Uncertainty estimation in deep 2d echocardiography segmentation, arXiv preprint arXiv:2005.09349 (2020).

[323] W. Zhu, A. Myronenko, Z. Xu, W. Li, H. Roth, Y. Huang, F. Milletari, D. Xu, Neurreg: Neural registration and its application to image segmentation, in: The IEEE Winter Conference on Applications of Computer Vision, 2020, pp. 3617–3626.

[324] C. Bian, C. Yuan, J. Wang, M. Li, X. Yang, S. Yu, K. Ma, J. Yuan, Y. Zheng, Uncertainty-aware domain alignment for anatomical structure segmentation, Medical Image Analysis (2020) 101732.

[325] S. A. Kohl, B. Romera-Paredes, K. H. Maier-Hein, D. J. Rezende, S. Eslami, P. Kohli, A. Zisserman, O. Ronneberger, A hierarchical probabilistic u-net for modeling multi-scale ambiguities, arXiv preprint arXiv:1905.13077 (2019).

[326] M. Yin, A. Yazdani, G. E. Karniadakis, One-dimensional modeling of fractional flow reserve in coronary artery disease: Uncertainty quantification and bayesian optimization, Computer Methods in Applied Mechanics and Engineering 353 (2019) 66–85.

[327] B. Li, T. S. Alstrøm, On uncertainty estimation in active learning for image segmentation, arXiv preprint arXiv:2007.06364 (2020).

[328] N. Strodthoff, P. Wagner, T. Schaeffter, W. Samek, Deep learning for ecg analysis: Benchmarks and insights from ptb-xl, arXiv preprint arXiv:2004.13701 (2020).

[329] C. Stoean, R. Stoean, M. Atencia, M. Abdar, L. Velázquez-Pérez, A. Khosravi, S. Nahavandi, U. R. Acharya, G. Joya, Automated detection of presymptomatic conditions in spinocerebellar ataxia type 2 using monte carlo dropout and deep neural network techniques with electrooculogram signals, Sensors 20 (11) (2020) 3032.

[330] A. Koulouri, V. Rimpiläinen, Simultaneous skull conductivity and focal source imaging from eeg recordings with the help of bayesian uncertainty modelling, in: European Medical and Biological Engineering Conference, Springer, 2020, pp. 1019–1027.

[331] J.-T. Chien, Y.-C. Ku, Bayesian recurrent neural network for language modeling, IEEE transactions on neural networks and learning systems 27 (2) (2015) 361–374.

[332] V. Vincze, Uncertainty detection in natural language texts, Ph.D. thesis, szte (2015).

[333] Y. Chen, T. A. Lasko, Q. Mei, J. C. Denny, H. Xu, A study of active learning methods for named entity recognition in clinical text, Journal of biomedical informatics 58 (2015) 11–18.

[334] L. Kong, G. Melis, W. Ling, L. Yu, D. Yogatama, Variational smoothing in recurrent neural network language models, arXiv preprint arXiv:1901.09296 (2019).

[335] X. Han, B. Li, Z. Wang, An attention-based neural framework for uncertainty identification on social media texts, Tsinghua Science and Technology 25 (1) (2019) 117–126.

[336] Z. Zhang, S. Zohren, S. Roberts, Bdlob: Bayesian deep convolutional neural networks for limit order books, arXiv preprint arXiv:1811.10041 (2018).

[337] T. Z. Xiao, A. N. Gomez, Y. Gal, Wat heb je gezegd? detecting out-of-distribution translations with variational transformers, in: Third workshop on Bayesian Deep Learning (NeurIPS 2018), 2019.

[338] M. Ott, M. Auli, D. Grangier, M. Ranzato, Analyzing uncertainty in neural machine translation, arXiv preprint arXiv:1803.00047 (2018).

[339] J. Vaicenavicius, D. Widmann, C. Andersson, F. Lindsten, J. Roll, T. B. Schön, Evaluating model calibration in classification, arXiv preprint arXiv:1902.06977 (2019).

[340] D. Tran, M. W. Dusenberry, M. van der Wilk, D. Hafner, Bayesian layers: A module for neural network uncertainty, arXiv preprint arXiv:1812.03973 (2018).

[341] J. Mena, O. Pujol, J. Vitrià, Dirichlet uncertainty wrappers for actionable algorithm accuracy accountability and auditability, arXiv preprint arXiv:1912.12628 (2019).

[342] Y. Xiao, W. Y. Wang, Quantifying uncertainties in natural language processing tasks, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 33, 2019, pp. 7322–7329.

[343] X. Zhang, S. Mahadevan, Bayesian neural networks for flight trajectory prediction and safety assessment, Decision Support Systems 131 (2020) 113246.

[344] T. Vandal, E. Kodra, J. Dy, S. Ganguly, R. Nemani, A. R. Ganguly, Quantifying uncertainty in discrete-continuous and skewed data with bayesian deep learning, in: Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2018, pp. 2377–2386.

[345] E. Nalisnick, J. M. Hernández-Lobato, Automatic depth determination for bayesian resnets, in: Third workshop on Bayesian Deep Learning (NeurIPS 2018), 2018.

[346] B. Bischke, P. Helber, J. Folz, D. Borth, A. Dengel, Multi-task learning for segmentation of building footprints with deep neural networks, in: 2019 IEEE International Conference on Image Processing (ICIP), IEEE, 2019, pp. 1480–1484.

[347] G. P. Meyer, N. Thakurdesai, Learning an uncertainty-aware object detector for autonomous driving, arXiv preprint arXiv:1910.11375 (2019).

[348] P. L. McDermott, C. K. Wikle, Bayesian recurrent neural network models for forecasting and quantifying uncertainty in spatial-temporal data, Entropy 21 (2) (2019) 184.

[349] J. Yu, M. W. Lam, S. Hu, X. Wu, X. Li, Y. Cao, X. Liu, H. Meng, Comparative study of parametric and representation uncertainty modeling for recurrent neural network language models., in: INTERSPEECH, 2019, pp. 3510–3514.

[350] N. Geneva, N. Zabaras, Quantifying model form uncertainty in reynolds-averaged turbulence models with bayesian deep neural networks, Journal of Computational Physics 383 (2019) 125–147.

[351] L. Feng, S. Xu, F. Wang, S. Liu, H. Qiao, Rough extreme learning machine: A new classification method based on uncertainty measure, Neurocomputing 325 (2019) 269–282.

[352] M. Walmsley, L. Smith, C. Lintott, Y. Gal, S. Bamford, H. Dickinson, L. Fortson, S. Kruk, K. Masters, C. Scarlata, et al., Galaxy zoo: probabilistic morphology through bayesian cnns and active learning, Monthly Notices of the Royal Astronomical Society 491 (2) (2020) 1554–1574.

[353] M. Henaff, A. Canziani, Y. LeCun, Model-predictive policy learning with uncertainty regularization for driving in dense traffic, arXiv preprint arXiv:1901.02705 (2019).

[354] Y. Zhang, et al., Bayesian semi-supervised learning for uncertainty-calibrated prediction of molecular properties and active learning, Chemical Science 10 (35) (2019) 8154–8163.

[355] D. Ruhe, G. Cina, M. Tonutti, D. de Bruin, P. Elbers, Bayesian modelling in practice: Using uncertainty to improve trustworthiness in medical applications, arXiv preprint arXiv:1906.08619 (2019).

[356] N. Neverova, D. Novotny, A. Vedaldi, Correlated uncertainty for learning dense correspondences from noisy labels, in: Advances in Neural Information Processing Systems, 2019, pp. 920–928.

[357] Y. Gal, J. Hron, A. Kendall, Concrete dropout, in: Advances in neural information processing systems, 2017, pp. 3581–3590.

[358] F. Verdoja, J. Lundell, V. Kyrki, Deep network uncertainty maps for indoor navigation, in: 2019 IEEE-RAS 19th International Conference on Humanoid Robots (Humanoids), IEEE, 2019, pp. 112–119.

[359] N. G. Polson, V. Sokolov, et al., Deep learning: A bayesian perspective, Bayesian Analysis 12 (4) (2017) 1275–1304.

[360] A. Brusaferri, M. Matteucci, P. Portolani, A. Vitali, Bayesian deep learning based method for probabilistic forecast of day-ahead electricity prices, Applied Energy 250 (2019) 1158–1175.

[361] M. Kampffmeyer, A.-B. Salberg, R. Jenssen, Semantic segmentation of small objects and modeling of uncertainty in urban remote sensing images using deep convolutional neural networks, in: Proceedings of the IEEE conference on computer vision and pattern recognition workshops, 2016, pp. 1–9.

[362] M. Sun, T. Zhang, Y. Wang, G. Strbac, C. Kang, Using bayesian deep learning to capture uncertainty for residential net load forecasting, IEEE Transactions on Power Systems 35 (1) (2019) 188–201.

[363] Y. Zeldes, S. Theodorakis, E. Solodnik, A. Rotman, G. Chamiel, D. Friedman, Deep density networks and uncertainty in recommender systems, arXiv preprint arXiv:1711.02487 (2017).

[364] L. Sun, J. Guo, Y. Zhu, A multi-aspect user-interest model based on sentiment analysis and uncertainty theory for recommender systems, Electronic Commerce Research (2018) 1–26.

[365] K. Jasberg, S. Sizov, Bayesian brain meets bayesian recommender-towards systems with empathy for the human nature, arXiv preprint arXiv:1706.08319 (2017).

[366] B. Xu, Z. Chen, Multi-level fusion based 3d object detection from monocular images, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 2345–2353.

[367] H. Li, X.-J. Wu, J. Kittler, Infrared and visible image fusion using a deep learning framework, in: 2018 24th International Conference on Pattern Recognition (ICPR), IEEE, 2018, pp. 2705–2710.

[368] S. Nemati, R. Rohani, M. E. Basiri, M. Abdar, N. Y. Yen, V. Makarenkov, A hybrid latent space data fusion method for multimodal emotion recognition, IEEE Access 7 (2019) 172948–172964.

[369] K. Tian, Y. Xu, S. Zhou, J. Guan, Versatile multiple choice learning and its application to vision computing, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 6349–6357.

[370] A. Wasay, B. Hentschel, Y. Liao, S. Chen, S. Idreos, Mothernets: Rapid deep ensemble learning, in: Proceedings of the 3rd MLSys Conference (MLSys), 2020.

[371] K. Lee, C. Hwang, K. Park, J. Shin, Confident multiple choice learning, arXiv preprint arXiv:1706.03475 (2017).

[372] W. H. Beluch, T. Genewein, A. Nürnberger, J. M. Köhler, The power of ensembles for active learning in image classification, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 9368–9377.

[373] Y. Tan, B. Jin, X. Yue, Y. Chen, A. S. Vincentelli, Exploiting uncertainties from ensemble learners to improve decision-making in healthcare ai, arXiv preprint arXiv:2007.06063 (2020).

[374] H. J. Lee, S. T. Kim, N. Navab, Y. M. Ro, Efficient ensemble model generation for uncertainty estimation with bayesian approximation in segmentation, arXiv preprint arXiv:2005.10754 (2020).

[375] L. Tran, B. S. Veeling, K. Roth, J. Swiatkowski, J. V. Dillon, J. Snoek, S. Mandt, T. Salimans, S. Nowozin, R. Jenatton, Hydra: Preserving ensemble diversity for model distillation, arXiv preprint arXiv:2001.04694 (2020).

[376] J. Caldeira, B. Nord, Deeply uncertain: Comparing methods of uncertainty quantification in deep learning algorithms, Machine Learning: Science and Technology 2 (1) (2020) 1–9.

[377] Y. Yao, An outline of a theory of three-way decisions, in: International Conference on Rough Sets and Current Trends in Computing, Springer, 2012, pp. 1–17.

[378] Y. Ben-Haim, Info-gap decision theory: decisions under severe uncertainty, Elsevier, 2006.

[379] V. A. Marchau, W. E. Walker, P. J. Bloemen, S. W. Popper, Decision making under deep uncertainty: From theory to practice, Springer Nature, 2019.

[380] J. Gordon, J. Bronskill, M. Bauer, S. Nowozin, R. E. Turner, Meta-learning probabilistic inference for prediction, arXiv preprint arXiv:1805.09921 (2018).

[381] H. B. Lee, H. Lee, D. Na, S. Kim, M. Park, E. Yang, S. J. Hwang, Learning to balance: Bayesian meta-learning for imbalanced and out-of-distribution tasks, arXiv preprint arXiv:1905.12917 (2019).

[382] C. Nguyen, T.-T. Do, G. Carneiro, Uncertainty in model-agnostic meta-learning using variational inference, in: The IEEE Winter Conference on Applications of Computer Vision, 2020, pp. 3090–3100.

[383] S. X. Hu, P. G. Moreno, Y. Xiao, X. Shen, G. Obozinski, N. D. Lawrence, A. Damianou, Empirical bayes transductive meta-learning with synthetic gradients, arXiv preprint arXiv:2004.12696 (2020).

[384] M. Qu, T. Gao, L.-P. Xhonneux, J. Tang, Few-shot relation extraction via bayesian meta-learning on relation graphs, in: International Conference on Machine Learning, PMLR, 2020, pp. 7867–7876.

[385] S. Sinha, S. Ebrahimi, T. Darrell, Variational adversarial active learning, in: Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 5972–5981.

[386] S. Zaidi, A. Zela, T. Elsken, C. Holmes, F. Hutter, Y. W. Teh, Neural ensemble search for performant and calibrated predictions, arXiv preprint arXiv:2006.08573 (2020).

[387] D. Hendrycks, K. Lee, M. Mazeika, Using pre-training can improve model robustness and uncertainty, arXiv preprint arXiv:1901.09960 (2019).

[388] W. Van Gansbeke, S. Vandenhende, S. Georgoulis, M. Proesmans, L. Van Gool, Learning to classify images without labels, arXiv preprint arXiv:2005.12320 (2020).

[389] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, in: Advances in neural information processing systems, 2017, pp. 5998–6008.

[390] Q. Wang, B. Wu, P. Zhu, P. Li, W. Zuo, Q. Hu, Eca-net: Efficient channel attention for deep convolutional neural networks, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 11534–11542.

[391] D. Hendrycks, M. Mazeika, T. Dietterich, Deep anomaly detection with outlier exposure, arXiv preprint arXiv:1812.04606 (2018).

[392] S. Yun, D. Han, S. J. Oh, S. Chun, J. Choe, Y. Yoo, Cutmix: Regularization strategy to train strong classifiers with localizable features, in: Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 6023–6032.

[393] J. Ren, P. J. Liu, E. Fertig, J. Snoek, R. Poplin, M. Depristo, J. Dillon, B. Lakshminarayanan, Likelihood ratios for out-of-distribution detection, in: Advances in Neural Information Processing Systems, 2019, pp. 14707–14718.

[394] S. Padhy, Z. Nado, J. Ren, J. Liu, J. Snoek, B. Lakshminarayanan, Revisiting one-vs-all classifiers for predictive uncertainty and out-of-distribution detection in neural networks, arXiv preprint arXiv:2007.05134 (2020).

[395] L. Kong, J. Sun, C. Zhang, Sde-net: Equipping deep neural networks with uncertainty estimates, arXiv preprint arXiv:2008.10546 (2020).

[396] J. Postels, H. Blum, C. Cadena, R. Siegwart, L. Van Gool, F. Tombari, Quantifying aleatoric and epistemic uncertainty using density estimation in latent space, arXiv preprint arXiv:2012.03082 (2020).

[397] D. Ulmer, G. Cinà, Know your limits: Monotonicity & softmax make neural classifiers overconfident on ood data, arXiv preprint arXiv:2012.05329 (2020).

[398] D. Ha, A. Dai, Q. V. Le, Hypernetworks, arXiv preprint arXiv:1609.09106 (2016).

[399] A. Navon, A. Shamsian, G. Chechik, E. Fetaya, Learning the pareto front with hypernetworks, arXiv preprint arXiv:2010.04104 (2020).

[400] T. Galanti, L. Wolf, On the modularity of hypernetworks, Advances in Neural Information Processing Systems 33 (2020).

[401] D. Krueger, C.-W. Huang, R. Islam, R. Turner, A. Lacoste, A. Courville, Bayesian hypernetworks, arXiv preprint arXiv:1710.04759 (2017).

[402] C. Zeno, I. Golan, E. Hoffer, D. Soudry, Task agnostic continual learning using online variational bayes, arXiv preprint arXiv:1803.10123 (2018).

[403] S. Farquhar, Y. Gal, A unifying bayesian view of continual learning, arXiv preprint arXiv:1902.06494 (2019).

[404] S. Kessler, V. Nguyen, S. Zohren, S. Roberts, Indian buffet neural networks for continual learning, arXiv preprint arXiv:1912.02290 (2019).

[405] Y. Chen, T. Diethe, N. Lawrence, Facilitating bayesian continual learning by natural gradients and stein gradients, arXiv preprint arXiv:1904.10644 (2019).

[406] A. Kumar, S. Chatterjee, P. Rai, Bayesian structure adaptation for continual learning, arXiv (2019) arXiv–1912.

[407] H. Li, P. Barnaghi, S. Enshaeifar, F. Ganz, Continual learning using bayesian neural networks, IEEE Transactions on Neural Networks and Learning Systems (2020).

[408] P. Pan, S. Swaroop, A. Immer, R. Eschenhagen, R. E. Turner, M. E. Khan, Continual deep learning by functional regularisation of memorable past, arXiv preprint arXiv:2004.14070 (2020).

[409] J. Zhou, G. Cui, Z. Zhang, C. Yang, Z. Liu, L. Wang, C. Li, M. Sun, Graph neural networks: A review of methods and applications, arXiv preprint arXiv:1812.08434 (2018).

[410] P.-E. Sarlin, D. DeTone, T. Malisiewicz, A. Rabinovich, Superglue: Learning feature matching with graph neural networks, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 4938–4947.

[411] F. d. A. Belbute-Peres, T. Economon, Z. Kolter, Combining differentiable pde solvers and graph neural networks for fluid flow prediction, in: International Conference on Machine Learning, PMLR, 2020, pp. 2402–2411.

[412] D. Luo, W. Cheng, D. Xu, W. Yu, B. Zong, H. Chen, X. Zhang, Parameterized explainer for graph neural network, Advances in Neural Information Processing Systems 33 (2020).

[413] W. Shi, R. Rajkumar, Point-gnn: Graph neural network for 3d object detection in a point cloud, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 1711–1719.

[414] A. Hasanzadeh, E. Hajiramezanali, S. Boluki, M. Zhou, N. Duffield, K. Narayanan, X. Qian, Bayesian graph neural networks with adaptive connection sampling, arXiv preprint arXiv:2006.04064 (2020).

[415] F. Teimury, S. Pal, A. Amini, M. Coates, Estimation of time-series on graphs using bayesian graph convolutional neural networks, in: Wavelets and Sparsity XVIII, Vol. 11138, International Society for Optics and Photonics, 2019, p. 111380Y.

[416] P. Elinas, E. V. Bonilla, L. Tiao, Variational inference for graph convolutional networks in the absence of graph data and adversarial settings, Advances in Neural Information Processing Systems 33 (2020).

[417] J. Sun, W. Guo, D. Zhang, Y. Zhang, F. Regol, Y. Hu, H. Guo, R. Tang, H. Yuan, X. He, et al., A framework for recommending accurate and diverse items using bayesian graph convolutional neural networks, in: Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2020, pp. 2030–2039.

[418] L. Ma, J. Cui, B. Yang, Deep neural architecture search with deep graph bayesian optimization, in: 2019 IEEE/WIC/ACM International Conference on Web Intelligence (WI), IEEE, 2019, pp. 500–507.

[419] J. Snoek, H. Larochelle, R. P. Adams, Practical bayesian optimization of machine learning algorithms, Advances in neural information processing systems 25 (2012) 2951–2959.

[420] P. I. Frazier, A tutorial on bayesian optimization, arXiv preprint arXiv:1807.02811 (2018).

[421] J. Wu, S. Toscano-Palmerin, P. I. Frazier, A. G. Wilson, Practical multi-fidelity bayesian optimization for hyperparameter tuning, in: Uncertainty in Artificial Intelligence, PMLR, 2020, pp. 788–798.

[422] M. Balandat, B. Karrer, D. Jiang, S. Daulton, B. Letham, A. G. Wilson, E. Bakshy, Botorch: A framework for efficient monte-carlo bayesian optimization, Advances in Neural Information Processing Systems 33 (2020).

[423] F. Sorourifar, G. Makrygirgos, A. Mesbah, J. A. Paulson, A data-driven automatic tuning method for mpc under uncertainty using constrained bayesian optimization, arXiv preprint arXiv:2011.11841 (2020).

[424] A. Kumar, P. S. Liang, T. Ma, Verified uncertainty calibration, Advances in Neural Information Processing Systems 32 (2019) 3792–3803.

[425] C. Guo, G. Pleiss, Y. Sun, K. Q. Weinberger, On calibration of modern neural networks, arXiv preprint arXiv:1706.04599 (2017).

[426] K. Patel, W. Beluch, B. Yang, M. Pfeiffer, D. Zhang, Multi-class uncertainty calibration via mutual information maximization-based binning, arXiv preprint arXiv:2006.13092 (2020).

[427] X. Wang, M. Long, J. Wang, M. Jordan, Transferable calibration with lower bias and variance in domain adaptation, Advances in Neural Information Processing Systems 33 (2020).

[428] D. Widmann, F. Lindsten, D. Zachariah, Calibration tests in multi-class classification: A unifying framework, in: Advances in Neural Information Processing Systems, 2019, pp. 12257–12267.

[429] A. Rahimi, A. Shaban, C.-A. Cheng, R. Hartley, B. Boots, Intra order-preserving functions for calibration of multi-class neural networks, Advances in Neural Information Processing Systems 33 (2020).

[430] S. Utpala, P. Rai, Quantile regularization: Towards implicit calibration of regression models, arXiv preprint arXiv:2002.12860 (2020).

[431] S. Zhao, T. Ma, S. Ermon, Individual calibration with randomized forecasting, in: International Conference on Machine Learning, PMLR, 2020, pp. 11387–11397.

[432] M. Rawat, M. Wistuba, M.-I. Nicolae, Harnessing model uncertainty for detecting adversarial examples, in: NIPS Workshop on Bayesian Deep Learning, 2017.

[433] J. Su, M. Cvitkovic, F. Huang, Sampling-free learning of bayesian quantized neural networks, arXiv preprint arXiv:1912.02992 (2019).

[434] M. Haußmann, F. A. Hamprecht, M. Kandemir, Sampling-free variational inference of bayesian neural networks by variance backpropagation, in: Uncertainty in Artificial Intelligence, PMLR, 2020, pp. 563–573.

[435] Y. Shi, A. K. Jain, Probabilistic face embeddings, in: Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 6902–6911.

[436] T. R. Scott, K. Ridgeway, M. C. Mozer, Stochastic prototype embeddings, arXiv preprint arXiv:1909.11702 (2019).

[437] F. D. S. Ribeiro, G. Leontidis, S. D. Kollias, Capsule routing via variational bayes., in: AAAI, 2020, pp. 3749–3756.

[438] R. Loftin, M. E. Taylor, M. L. Littman, J. MacGlashan, B. Peng, D. L. Roberts, Open problems for online bayesian inference in neural networks, in: Bayesian Deep Learning Workshop at NeurIPS, 2016.

[439] H. Ahn, S. Cha, D. Lee, T. Moon, Uncertainty-based continual learning with adaptive regularization, in: Advances in Neural Information Processing Systems, 2019, pp. 4394–4404.

[440] C. V. Nguyen, Y. Li, T. D. Bui, R. E. Turner, Variational continual learning, arXiv preprint arXiv:1710.10628 (2017).

[441] H. Tseran, M. E. Khan, T. Harada, T. D. Bui, Natural variational continual learning, in: Continual Learning Workshop@ NeurIPS, Vol. 2, 2018.

[442] A. Lacoste, B. Oreshkin, W. Chung, T. Boquet, N. Rostamzadeh, D. Krueger, Uncertainty in multitask transfer learning, arXiv preprint arXiv:1806.07528 (2018).

[443] T. A. Nguyen, H. Jeong, E. Yang, S. J. Hwang, Clinical risk prediction with temporal probabilistic asymmetric multi-task learning, arXiv preprint arXiv:2006.12777 (2020).

[444] S. Wang, Y. Liu, C. Wang, H. Luan, M. Sun, Improving back-translation with uncertainty-based confidence estimation, arXiv preprint arXiv:1909.00157 (2019).

[445] K. Hama, T. Matsubara, K. Uehara, J. Cai, Exploring uncertainty measures for image-caption embedding-and-retrieval task, arXiv preprint arXiv:1904.08504 (2019).

[446] L. Zhou, H. Palangi, L. Zhang, H. Hu, J. J. Corso, J. Gao, Unified vision-language pre-training for image captioning and vqa., in: AAAI, 2020, pp. 13041–13049.

[447] M. Cornia, M. Stefanini, L. Baraldi, R. Cucchiara, Meshed-memory transformer for image captioning, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 10578–10587.

[448] Y. Pan, T. Yao, Y. Li, T. Mei, X-linear attention networks for image captioning, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 10971–10980.

[449] S. Antol, A. Agrawal, J. Lu, M. Mitchell, D. Batra, C. Lawrence Zitnick, D. Parikh, Vqa: Visual question answering, in: Proceedings of the IEEE international conference on computer vision, 2015, pp. 2425–2433.

[450] J. Lu, J. Yang, D. Batra, D. Parikh, Hierarchical question-image co-attention for visual question answering, in: Advances in neural information processing systems, 2016, pp. 289–297.

[451] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, L. Zhang, Bottom-up and top-down attention for image captioning and visual question answering, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 6077–6086.

[452] Y. Goyal, T. Khot, D. Summers-Stay, D. Batra, D. Parikh, Making the v in vqa matter: Elevating the role of image understanding in visual question answering, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 6904–6913.

[453] Z. Yu, J. Yu, Y. Cui, D. Tao, Q. Tian, Deep modular co-attention networks for visual question answering, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2019, pp. 6281–6290.

[454] A. Wu, S. Nowozin, E. Meeds, R. E. Turner, J. M. Hernández-Lobato, A. L. Gaunt, Deterministic variational inference for robust bayesian neural networks, arXiv preprint arXiv:1810.03958 (2018).

[455] T. E. Boult, S. Cruz, A. R. Dhamija, M. Gunther, J. Henrydoss, W. J. Scheirer, Learning and the unknown: Surveying steps toward open world recognition, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 33, 2019, pp. 9801–9807.

[456] S. Ghosh, F. M. Delle Fave, J. S. Yedidia, Assumed density filtering methods for learning bayesian neural networks., in: AAAI, 2016, pp. 1589–1595.

[457] X. Lu, B. Van Roy, Ensemble sampling, in: Advances in neural information processing systems, 2017, pp. 3258–3266.

[458] M. Karamanis, F. Beutler, Ensemble slice sampling, arXiv preprint arXiv:2002.06212 (2020).

[459] F. Pourpanah, C. J. Tan, C. P. Lim, J. Mohamad-Saleh, A q-learning-based multi-agent system for data classification, Applied Soft Computing 52 (2017) 519 – 531.

[460] F. Pourpanah, R. Wang, C. P. Lim, X. Wang, M. Seera, C. J. Tan, An improved fuzzy artmap and q-learning agent model for pattern classification, Neurocomputing 359 (2019) 139 – 152.

Highlights

- Provided an extensive review of uncertainty quantification methods in deep learning.
- Covered most applied Bayesian approaches for uncertainty quantification.
- Listed notable ensemble techniques for quantifying uncertainty.
- Discussed various applications of uncertainty quantification methods.
- Summarized major open challenges and research gaps in uncertainty quantification.

Moloud Abdar: Conception of the project, design of methodology, paper collection, taxonomy design, summary of articles, data extraction from reviewed articles, drafting the manuscript, result discussion, revising the manuscript, editing. **Farhad Pourpanah:** Summary of articles, drafting the manuscript, result discussion, revising the manuscript, editing. **Sadiq Hussain:** Summary of articles, drafting the manuscript, result discussion, revising the manuscript. **Dana Rezazadegan:** Drafting the manuscript, revising the manuscript, data extraction from reviewed articles, editing. **Li Liu:** Result discussion, revising the manuscript, supervision. **Mohammad Ghavamzadeh:** Result discussion, revising the manuscript, supervision. **Paul Fieguth:** Result discussion, revising the manuscript, supervision. **Xiaochun Cao:** Result discussion, revising the manuscript, supervision. **Abbas Khosravi:** Result discussion, revising the manuscript, supervision. **U. Rajendra Acharya:** Result discussion, revising the manuscript, supervision. **Vladimir Makarenkov:** Result discussion, revising the manuscript, supervision. **Saeid Nahavandi:** Result discussion, revising the manuscript, supervision.

Conflict of Interest

The authors declare that they have no conflict of interest.