The following publication L. Xue, S. -H. Chung, L. Yang, X. -L. Wang and X. Zhang, "A Unified Uncertainty-Informed Approach for Risk Management of Deep Learning Models in the Open World," in IEEE Transactions on Emerging Topics in Computational Intelligence, vol. 10, no. 2, pp. 2121-2135, April 2026 is available at https://doi.org/10.1109/TETCI.2025.3647582.

A unified uncertainty-informed approach for risk management of deep learning models in the open world

Long Xue, Sai-Ho Chung, Lechang Yang, Xiao-Lin Wang, and Xiaoge Zhang, *Senior Member, IEEE*

***Abstract*—Equipping deep learning models with a principled uncertainty quantification (UQ) has become essential to ensure their reliable performance in open-world environments. To address uncertainty arising from two prevalent sources - distribution shifts and out-of-distribution (OOD) inputs, this paper presents a unified, uncertainty-informed approach for quantifying and managing the risks these factors pose to deep learning models. Toward this goal, we leverage a principled UQ approach, Spectral-normalized Neural Gaussian Process (SNGP), to quantify the epistemic uncertainty associated with model predictions. Unlike other UQ methods in the literature, SNGP offers two distinctive properties: (1) spectral normalization applied to hidden layer weights to preserve relative distances among data points throughout feature transformations, and (2) replacement of the output layer with a Gaussian process to produce distance-aware uncertainty estimates. Using the uncertainty estimates from SNGP, we employ Youden's index to derive an optimal threshold that categorizes predictions into different risk levels, enabling uncertainty-informed decision making. Experiments on two datasets of varying scale demonstrate that the proposed method facilitates effective risk assessment and management in open-world settings. Computational results show that the proposed method achieves predictive performance comparable to Monte Carlo dropout and deep ensembles, while providing more computationally efficient, consistent, and principled uncertainty estimates under no shift, distribution shift, and OOD conditions.**

***Index Terms*—Deep learning; Uncertainty quantification; Distribution shift; Out-of-distribution; Uncertainty-informed risk management**

I. Introduction

DEEP neural networks (DNNs) have been increasingly applied across a wide spectrum of domains, including object recognition [1, 2], chatbots [3], smart cities [4], and many others. While DNNs have achieved outstanding performance in various real-world applications, their translation and deployment in safety-critical applications remain strikingly low due to the lack of mature risk assessment and management of artificial intelligence (AI) models under open environments [5–8]. As revealed by several studies in the literature [9, 10], DNNs tend to be overconfident in the sense that they produce unreasonably high confidence for out-of-distribution (OOD) inputs. This overconfidence is especially concerning when the distribution of unseen testing data significantly differs from that of the training data. This situation occurs frequently and unavoidably after the model is deployed in an open-world setting [11, 12]. Since a misleading prediction might endanger the safety of users in high-stakes applications such as medical diagnosis and autonomous driving [13, 14], it is important to equip DNN models with a rigorous uncertainty quantification (UQ) to effectively assess, mitigate, and manage potential adverse outcomes [15, 16]. Such research and development efforts facilitate building reliable and safe AI models in the long term.

In the risk analysis field, uncertainty plays a pivotal role in defining, modeling, assessing, and managing the risk across various engineering systems [17, 18]. By definition, uncertainty refers to the lack of knowledge in estimating a hypothesis, a quantity, or the occurrence of an event [19]. Risk analysts often utilize probability, a generic modeling theory and a useful mathematical tool, to characterize uncertainty [20–22]. A range of approaches grounded in probability theory have been developed to assess and analyze the uncertainty, such as sampling-based methods [23] and the use of probability distributions [20, 24]. In addition, Bayesian subjective methods, interval probabilities, fuzzy theory, and qualitative approaches have also been developed [25]. These methods have been applied in a wide range of domains, such as engineering and infrastructure, occupation health and safety, and ecological risk assessment [18, 26, 27]. Nevertheless, these methods traditionally used for risk analysis in conventional engineering systems are not well-suited for measuring and assessing uncertainty in the context of DNNs for the following reasons: (I) Predictability and explainability [28]: DNN models lack transparency and interpretability, and they work like a black box. This will complicate the uncertainty modeling and risk management. (II) Data dependency: DNNs inherently rely on the data they are trained on. If the data is biased or defective, it will impair the predictions produced by the model. Data dependency might lead the model to create potentially dangerous situations and undesirable outcomes in high-stakes decision settings—a challenge that is rarely encountered in the risk analysis for traditional engineering systems. (III) Relearning and adaptability: DNNs can relearn as new data comes in. This relearning can introduce unknown uncertainties since DNN's behavior may change over time. To the best of our knowledge, limited research has been conducted to investigate DNN's uncertainty, risk assessment, and risk management when DNN

L. Xue, S. Chung, and X. Zhang are with the Department of Industrial and Systems Engineering, The Hong Kong Polytechnic University, Kowloon, Hong Kong SAR, China.

L. Yang is with the School of Mechanical Engineering, University of Science and Technology Beijing, Beijing 100083, China.

X. Wang is with the Business School, Sichuan University, Chengdu 610065, China.

Correspondence to: Room EF622, Department of Industrial and Systems Engineering, The Hong Kong Polytechnic University, 11 Yuk Choi Rd, Hung Hom, Hong Kong. Email: xiaoge.zhang@polyu.edu.hk.

models face distributional shifts in the open environment.

Accurate and reliable uncertainty estimation provides an informative signal that helps end users determine when a DNN's predictions can or cannot be trusted. Importantly, by leveraging predictive uncertainty, the DNN gains the ability to express "I don't know" by indicating the confidence level associated with its predictions. A higher uncertainty estimate naturally reflects a greater likelihood of prediction failure, and vice versa. By quantifying model prediction uncertainty in a given scenario and understanding the potential consequences, decision-makers gain a nuanced view of the risks associated with relying on the model's output. For example, when a self-driving car encounters a previously unseen situation, a high uncertainty estimate can alert the driver that the model's prediction is unreliable, prompting timely human intervention. Therefore, developing principled and robust UQ methods is essential for accurately characterizing uncertainty, supporting risk assessment, and enabling risk-informed decision-making in safety-critical applications.

This paper examines two major uncertainty sources in open-world environments: distribution shifts and OOD inputs. Typically, the training, validation, and testing data are assumed to be independently and identically distributed (i.i.d) samples drawn from the same distribution. However, this assumption does not always hold, when the model is deployed in a constantly changing environment. Over the past few years, distributional shifts have received increasing attention due to the pervasive and unpredictable variations that occur in real-world environments [29–31]. By definition, distributional shift refers to the mismatch between the training data (seen data) and the testing data (unseen data) [32, 33]. In other words, the training data distribution differs from the testing data distribution to some extent [30]. For example, consider a model developed to predict the risk of obesity based on the dataset from a specific population - old white males living in rural areas. If this model is applied to a different population, such as young Asian females living in urban areas, it performance may degrade significantly. This is because the distribution of the underlying data has changed - the training data features certain characteristics (old, white, male, rural), but the testing data has different characteristics (young, Asian, female, urban). In this case, the testing data is viewed as an instance of the distribution shift. If the distribution shift is ignored, the resultant model performance will degrade largely, thus increasing consequential health risks (e.g., false positives). In healthcare, such issue is particularly concerning as models are often developed using data from a specific subgroup of patients and then applied to broader patient population [34, 35]. Besides distribution shift, OOD data also presents a significant challenge to the trustworthiness and reliability of DNN applications in high-stakes domains. When confronting OOD inputs that fall outside the scope of the training distribution, DNN models often exhibit a substantial performance drop. For example, in the context of self-driving, OOD examples arise when self-driving cars encounter never-seen road scenarios that are not represented in the training data. Consider an autonomous driving system developed using data collected in California, where the weather is predominantly sunny. The developed model performs well since it has learned to recognize and respond to various traffic situations in California, such as traffic signs, pedestrians, vehicles, etc. However, if the same self-driving system is deployed on snowy roads in Alaska, it may struggle to identify traffic signs, as it was not trained on data from snowy conditions. In this context, the snowy weather conditions represents an OOD instance to the trained model. In this case, the self-driving car might misinterpret the snow-covered road signs. This situation poses a serious challenge because training self-driving systems in every possible driving condition is an impossible task. Ideally, in the case of OOD situations, to stay safe, the self-driving system is expected to abstain from making decisions by saying "I do not know". Therefore, it is important to assess and handle the OOD input to ensure the safety of self-driving vehicles.

Bayesian methods for DNN uncertainty estimation assume a predefined prior distribution over model parameters, which is updated with training data to obtain the posterior distribution via Bayes' rule [36–38]. However, since the exact posterior inference is computationally intractable, approximation methods have often been proposed to infer the posterior distributions, including variational inference [39], and Laplace approximation [40]. Although these approximation approaches have achieved decent performance on uncertainty estimation to some extent, their performance is highly dependent on the assumed prior and suffers from demanding computational cost compared to non-Bayesian methods [10]. Monte Carlo dropout (MC dropout) and deep ensembles are two common methods to quantify the prediction uncertainty of neural networks. In essence, MC dropout [41] estimates model uncertainty by performing multiple forward passes through a neural network with dropout enabled during model inference. As a result, it generates a distribution of predictions for uncertainty estimates. Unlike MC dropout, deep ensembles train multiple neural networks independently and then combine their predictions to approximate the predictive distribution. However, both approaches suffer from two fundamental deficiencies. First, both approaches require multiple forward passes for uncertainty estimation, which largely increases the computational cost compared to their deterministic counterparts [37]. Second, they often struggle with OOD detection by assigning high confidence to OOD samples [42]. As revealed in [43], MC dropout and deep ensemble estimate the uncertainty of a given sample based on its distance from the decision boundary separating different classes. This UQ behavior leads them to assign overconfident uncertainty estimates to OOD samples as they overlook the distance of the sample from the training data. Except for the weaknesses mentioned in the existing UQ methods, limited research has been performed to utilize the estimated uncertainty to combat distribution shifts and OOD at the same time. In 2017, Lakshminarayanan et al. [10] assessed the robustness of deep ensemble-based method for uncertainty estimation using OOD samples from unseen classes. In addition, Ovadia et al. [36] comprehensively evaluated six existing uncertainty estimation methods under dataset shift on classification tasks. Furthermore, Liu et al. [42] proposed a distance-aware approach to quantify the predictive uncertainty of the OOD. The experimental results demonstrate the distance-aware method generates principled predictive uncertainty.

In this work, our goal is to develop a unified, uncertainty-

informed approach for assessing and managing the risks posed by distribution shifts and OOD data in the open-world deployment of DNNs. To this end, we consider three evaluation scenarios—no shift, distribution shift, and OOD—and systematically assess the quality of predictive uncertainty produced by several UQ methods. Normally, the larger the distributional shift, the higher the predictive uncertainty. If the input sample is OOD, the estimated uncertainty should be way higher to alert decision-makers to give up relying on the DNN model's prediction. To achieve this desirable behavior in UQ, we leverage a principled UQ method, spectral-normalized Gaussian process (SNGP), due to its two properties in distance-preserving representation learning and distance-aware output layer [42, 44]. Specifically, as the input data passes through multiple transformations in the hidden layers, the distance-preserving property ensures the relative distances between data points in the input space are maintained in the hidden space. In other words, the distance $\|h(\boldsymbol{x}) - h(\boldsymbol{x}')\|_H$ in the latent space has a meaningful correspondence to the distance $\|\boldsymbol{x} - \boldsymbol{x}'\|_X$ in the input space, see Eq. (4). The distance-awareness property is achieved by replacing the conventional dense output layer of neural network with an approximate Gaussian process (GP). Consequently, the magnitude of the distance in the latent space is translated into the corresponding predictive variance (i.e. predictive uncertainty) accordingly.

We quantitatively compare the performance of the proposed method against MC dropout and deep ensembles across three levels of distributional shift. Furthermore, we establish an uncertainty-informed risk assessment and management strategy to safeguard the DNN's applications. Our contributions are summarized as follows,

1) We consider a real-world classification problem in three scenarios (i.e., no distribution shift, distributional shift, and OOD) to mimic the potentially varying levels of open-world novelty. This setup allows us to illustrate the emerging risks that DNN models commonly face in practice.
2) We compare the performance of three UQ methods regarding the quality of UQ estimations and classification accuracy. A quantitative metric is introduced to assess the correlation between the estimated uncertainty and the degree of distributional shift. We empirically demonstrate that the proposed method not only achieves comparable predictive accuracy, but produces more consistent and reliable uncertainty estimations.
3) We reveal the high-quality UQ of the proposed method by uncovering the crucial role of distance preserving during feature transformation in uncertainty estimation. The Pearson correlation coefficient and the average predictive uncertainty are two metrics to evaluate the quality of estimated uncertainty across the three considered scenarios.
4) We develop an uncertainty-informed risk assessment and management strategy based on Younden's index to divide the uncertainty to quantify the risks in consideration of the specific scenario and its consequences. By establishing proper uncertainty thresholds, the uncertainty and its associated risk are well understood, and corresponding actions can be taken according to decision-makers' evaluations.

The rest of the paper is organized as follows. Section II summarizes the differences and similarities of uncertainty in the traditional risk analysis field and the context of DNNs. Section III briefly describes the two considered uncertainty sources in the open-world setting, introduces the underlying logic behind the SNGP approach in detail, and presents the uncertainty-informed risk management strategy. Section IV describes the datasets and MC dropout and deep ensemble approaches for computational study, and analyzes the performance of each method. Section IV-E explicates the utilization of uncertainty-informed risk assessment and management strategy based on Youden's index. Section V ends this paper with conclusions and future work.

II. Uncertainty modeling

Uncertainty has played a central role in the definition, quantification, and assessment of risk over the past decades [21, 22]. Measuring and quantifying uncertainty is vital for the sound risk assessment and management in high-stakes applications. We adopt the definition of uncertainty from The Society for Risk Analysis (SRA): "Imperfect or incomplete information/knowledge about a hypothesis, a quantity, or the occurrence of an event" [19]. Based on this definition, we elaborate on the similarities and differences of uncertainty from four aspects - uncertainty types, sources, modeling methods, and applications - in the traditional risk analysis domain and the uncertainty instantiations in the DNN context.

In risk analysis, uncertainty is typically categorized into epistemic and aleatoric (stochastic) uncertainty. In essence, epistemic uncertainty refers to "incomplete knowledge about a hypothesis, a quantify, and the occurrence of an event" [19] while aleatoric uncertainty arises from the "variation of quantities in a population of units (commonly represented/described by a probability model" [19]. To model these heterogeneous uncertainties, various methods have been developed in the literature. It is well-accepted that probability is the most fundamental theory used to measure and quantify the uncertainty in risk analysis. Following this, three different interpretations of probability - classical, frequentist, and subjective (judgment, knowledge-based) probabilities - have been applied to express and quantify the uncertainty [19, 21]. In addition to the probability methods, interval probability and fuzzy probability have also been adopted to model the uncertainty [18, 45].

In contrast, in the DNN field, uncertainty sources are often analyzed from the epistemic (model) and aleatoric (data) perspectives. Epistemic uncertainty arises from four primary sources: (i) uncertainty in the model structure, (ii) uncertainty introduced during the training procedure, (iii) variability in real-world environments such as distribution shifts, and (iv) uncertainty caused by unknown or OOD inputs [9]. Because DNNs depend heavily on the training distribution, they lack the knowledge to correctly interpret inputs that deviate from it. As a result, distributional shifts or OOD samples can lead to degraded or overconfident predictions. Data uncertainty refers to the inherent noise in the data and, unlike epistemic uncertainty, is irreducible. Owing to the complexity, data dependency, and adaptability mentioned earlier in Section I, existing treatments for uncertainty modeling in the context of traditional risk analysis are inapplicable to DNNs. In the context of DNNs, Bayesian, ensemble, and single network

TABLE I
COMPARISON OF **UNCERTAINTY MODELING** IN **TRADITIONAL RISK ANALYSIS AND DNN CONTEXTS**

Description	**Uncertainty modeling**		**Difference**
	In traditional risk analysis	In deep neural networks	
Category	Epistemic uncertainty, aleatoric (stochastic) uncertainty	Model (epistemic) uncertainty and data (aleatoric) uncertainty	No
Source	Epistemic uncertainty: "incomplete knowledge about a hypothesis, a quantity, or the occurrence of an event" [19]; Aleatoric uncertainty: "variation of quantities in a population of units (commonly represented/described by a probability model)" [19]	Epistemic (model) uncertainty: incomplete knowledge about the model (e.g., model structure and training procedure), variability in real world situations (e.g., distribution shift), and unknown data (e.g., out-of-distribution) Aleatoric (data) uncertainty: imprecise measurement system (e.g., imprecise sensors)	No
Quantitative method	Probabilistic analysis, fuzzy approach, probability interval, etc [45, 47]	Ensemble methods, Bayesian methods, single network deterministic methods, etc [9, 37, 48]	Yes
Application	Nuclear industry, occupation health and safety, infrastructure systems, security and defense, supply chain management, etc.	Medical image analysis, robotics, autonomous vehicle control, fraud detection, etc.	Yes

deterministic methods are common approaches adopted to measure and quantify the predictive uncertainty. In Bayesian methods, we obtain distributions over model parameters rather than point estimates, then propagate these to produce predictive distributions whose variance quantifies uncertainty. Since exact Bayesian inference is intractable, deep ensembles approximate this by training multiple identical architectures independently and using the variance of their aggregated predictions as uncertainty estimates. Alternatively, single-model deterministic methods achieve uncertainty quantification in one forward pass through techniques like distance preservation, GP, and prior networks [9, 14, 42, 46].

From an application standpoint, traditional risk analysis methods for UQ are primarily used in domains such as nuclear safety, engineering, infrastructure systems, and occupational health. In contrast, UQ methods developed for DNNs are typically applied to areas like medical image analysis, robotics, and autonomous systems. Table I provides a structured comparison of the major differences between these two domains in terms of uncertainty categories, sources, quantitative methods, and application areas.

III. PROPOSED METHODOLOGY

This section introduces the concepts of distribution shift and OOD inputs, followed by a detailed description of the SNGP method for UQ. It then explains how the resulting uncertainty estimates are used to build uncertainty-informed deep learning models and manage risks in open-world environments.

A. Distribution Shift and Out-of-Distribution

By definition, the distribution shift refers to the mismatch between the training and production data (i.e., test data) [33]. Fig. 1 illustrates the relationship between the training data, distribution shift, and OOD. As can be seen, the distribution of training data and distributional shifts are different in terms of the basic distribution characteristics, such as modes, distribution shape, variance, etc. Here, we view OOD as a completely different distribution from the training data distribution. Compared to the distribution shift, OOD is farther away from the training data distribution. In particular, the OOD does not overlap with the training data distribution. As reflected by the overlap between the training data and the distribution shift, the distribution shift only happens to some features and instances, but not all the features and instances.

DNN models generally perform the best on the non-shift distribution data (i.e., i.i.d) and the worst on the OOD. The

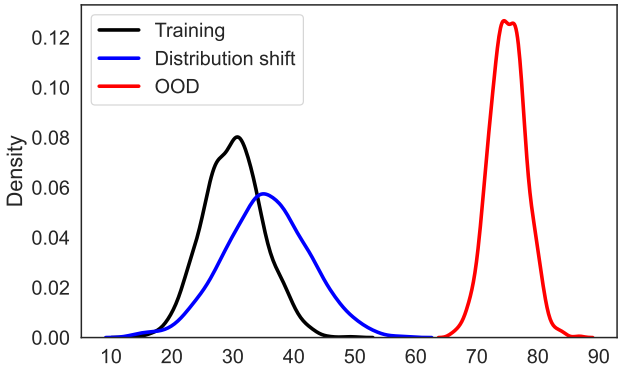


Fig. 1. A schematic diagram to illustrate **distribution shift and OOD**

larger the distributional shift, the poorer the model's prediction performance. Suppose inputs and target variables are denoted by $\boldsymbol{X} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_n\}$ and $\boldsymbol{Y} = \{y_1, y_2, \cdots, y_n\}$, where n denotes the number of data points, $\boldsymbol{x}_i$ $\left(\boldsymbol{x}_i \in \mathbb{R}^d, 1 \leq i \leq n\right)$ denotes the d-dimensional features associated with the i-th sample while y_i denotes the target corresponding to $\boldsymbol{x}_i$, and $\boldsymbol{x}_i \sim \mathcal{X}$, $y_i \sim \mathcal{Y}$. Mathematically, the independently and identically joint distribution of $\mathcal{X}$ and $\mathcal{Y}$ can be expressed as follows:

$$P_{\text{Test}}(\mathcal{X}, \mathcal{Y}) = P_{\text{Train}}(\mathcal{X}, \mathcal{Y}) \tag{1}$$

The distribution shift denotes the mismatch between the distributions of training data and the testing data. It can be mathematically defined as below [49]:

$$P_{\text{Test}}(\mathcal{X}) \neq P_{\text{Train}}(\mathcal{X}) \text{ and } P_{\text{Test}}(\mathcal{Y}) = P_{\text{Train}}(\mathcal{Y}) \tag{2}$$

For the OOD, we define it as follows:

$$P_{\text{Test}}(\mathcal{X}) \neq P_{\text{Train}}(\mathcal{X}) \text{ and } P_{\text{Test}}(\mathcal{Y}) \neq P_{\text{Train}}(\mathcal{Y}) \tag{3}$$

In OOD cases, where $P_{\text{Test}}(\mathcal{Y}) \neq P_{\text{Train}}(\mathcal{Y})$, the model has no knowledge of the true labels, as they lie outside the support of the training distribution. As a result, its predictions become unreliable and incorrect.

B. Uncertainty Quantification

We describe the SNGP method for UQ, highlighting its two key features that enable distance-preserving representations and distance-aware uncertainty estimation. SNGP seamlessly

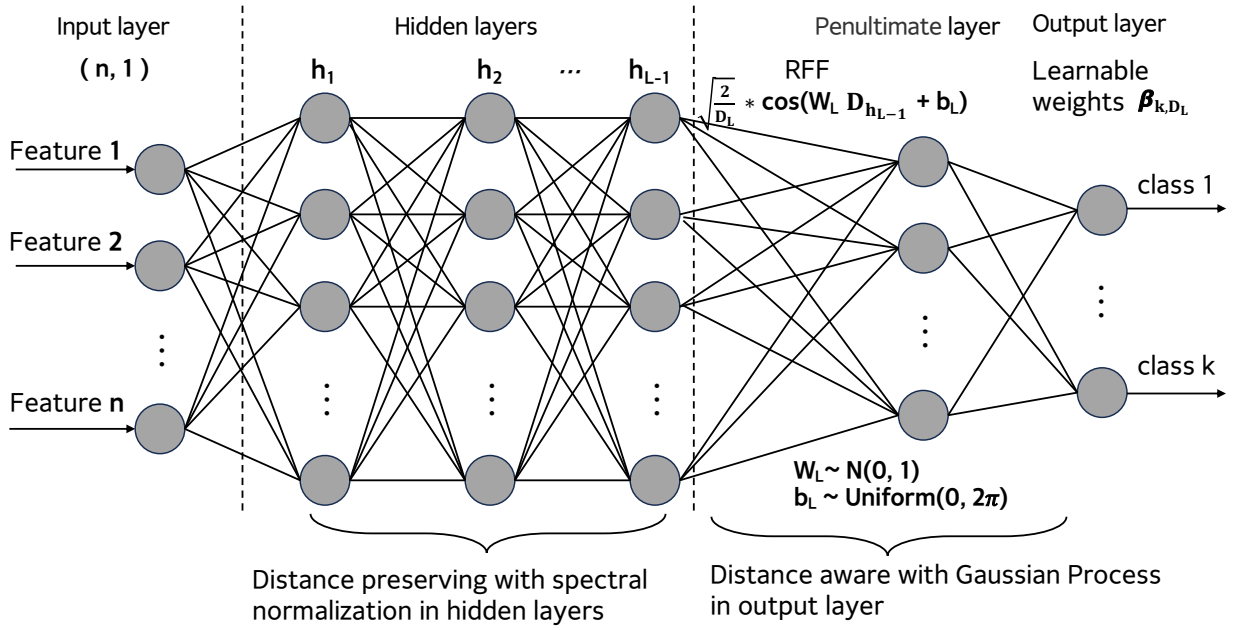


Fig. 2. A schematic illustration for **distance-aware uncertainty quantification** for DNN model

integrates with existing DNNs by combining the feature extraction capability of NNs with the distance-aware uncertainty estimates of GP. Architecturally, SNGP employs a standard NN as a feature extractor, followed by a Random Fourier Features (RFF)-based approximation to GP. This architecture enables SNGP not only to harness the powerful representation learning ability of NNs but also to provide an estimate of the uncertainty associated with the model predictions. Specifically, the uncertainty is captured as the predicted variance generated by the RFF-based GP layer. Importantly, the spectral normalization constraint ensures that feature representations learned by NNs are sensitive to input variations and capable of generalizing effectively due to the enforced smoothness [50]. As shown in Fig. 2, SNGP makes two critical changes to the regular neural network: adding spectral normalization on the weights of each hidden layer to make the feature mapping distance-preserving; replacing the traditional dense output layer with a GP to make the quantified uncertainty distance-sensitive. Note that RFF is adopted to approximate the GP kernel function in order to make the computation scalable [51]. These two prerequisites enable deep learning models to express uncertainty as a function of the distance between the training data and test samples in a principled way.

1) Distance Preserving with Spectral Normalization: Distance preserving entails that the distances between any pair of data points are maintained after they go through multiple transformations by the hidden layers in the DNN model, see $h_1, h_2, \cdots, h_{L-1}$ in Fig. 2. This enables the distances in the hidden space to have a meaningful correspondence to the distance in the input space. However, this distance-preserving property is usually not guaranteed for regular DNN models, which leads to a detrimental phenomenon known as feature collapse in neural network [52].

By applying spectral normalization over the weights of each hidden layer, the distance in the input space can be preserved in the hidden space appropriately. Consider a given input $\boldsymbol{x}$, the hidden mapping function $h(\boldsymbol{x})$, the output layer g, a deep learning classification model can simply be expressed as $logit(\boldsymbol{x}) = g \circ h(\boldsymbol{x})$. In this equation, the function $h(\boldsymbol{x})$ maps the input $\boldsymbol{x}$ into the representation $h(\boldsymbol{x})$ in the latent space, and the output layer g maps the $h(\boldsymbol{x})$ into the corresponding label. Given test samples $\boldsymbol{x}'$, the *bi-Lipschitz* condition in Eq. (4) needs to be satisfied to ensure that the distances in the hidden space $\|h(\boldsymbol{x}) - h(\boldsymbol{x}')\|_H$ have a strong correlation to the distances in the input space $\|\boldsymbol{x} - \boldsymbol{x}'\|_X$[42]. Mathematically, the mapping $h(\boldsymbol{x})$ should satisfy the *bi-Lipschitz* defined as follows [53]:

$$L_1 * \|\boldsymbol{x} - \boldsymbol{x}'\|_X \leq \|h(\boldsymbol{x}) - h(\boldsymbol{x}')\|_H \leq L_2 * \|\boldsymbol{x} - \boldsymbol{x}'\|_X \quad (4)$$

where L_1 and L_2 are positive lower and upper bounds ($0 < L_1 < 1 < L_2$) imposed on the Lipschitz constant of the feature extractor $h(\cdot)$, *i.e.*, $h(\boldsymbol{x})$ is the distance preserving function.

To this end, we apply spectral normalization over the weights of each hidden layer so that the distances in the hidden space have a corresponding relation to the distances in the input space. In spectral normalization, we set the spectral norm (i.e., the largest singular value) on each hidden weight matrix $\{W_l\}_{l=1}^{L-1}$ to less than 1, which is sufficient to ensure the hidden mapping $h(\boldsymbol{x})$ is distance preserving. We adopt the power iteration method to estimate the spectral norm $\hat{\lambda} \approx ||W_l||_2$ at every training step, and the spectral normalization on hidden layer weights W_l can be expressed as follows [54, 55]:

$$\mathbf{W}_l = \begin{cases} c * \mathbf{W}_l / \hat{\lambda} & \text{if } c < \hat{\lambda} \\ \mathbf{W}_l & \text{otherwise} \end{cases} \quad (5)$$

where $c > 0$ is employed to fine-tune the spectral upper bound on $||W_l||_2$ and to ensure $||W_l||_2 \leq c$.

2) Distance-Aware Uncertainty Estimation: To incorporate distance awareness into the output layer $g : h \to \mathcal{Y}$, the SNGP replaces the standard dense output layer with GP.

In the GP, the level of uncertainty at the test point $\boldsymbol{x}'$ is determined by its distance from the training data in the hidden space. Given N training samples $D = \{y_i, \mathbf{x_i}\}_{i=1}^{N}$ and the penultimate layer $h_i = h(\boldsymbol{x}_i)$, the output layer of the GP $\mathbf{g}_{N\times 1} = \big[g(h_1), \ldots, g(h_N)\big]^T$ is assumed to follow a multivariate normal prior distribution:

$$\begin{aligned} g_{N\times 1} &\sim \text{MVN}(\mathbf{0}_{N\times 1}, \boldsymbol{K}_{N\times N}) \\ \text{where } K_{i,j} &= \exp\left(-\frac{||h_i - h_j||_2^2}{2}\right) \end{aligned} \tag{6}$$

According to the Bayes' theorem, the posterior distribution is calculated as $p(g|D) \propto p(D|g) \cdot p(g)$. The $p(g)$ represents the prior of the GP described in Eq. (6). The $p(D|g)$ is the data likelihood indicating the probability of seeing the observed data D given the parameters g. However, inferring the exact GP posterior involves a high computational complexity of $O(N^3)$ due to the need to invert the $N \times N$ kernel matrix $\mathbf{K}$, which can be problematic when handling large datasets. As a consequence, we adopt the RFF method to approximate the kernel function[51]. The RFF-based approach results in a low-rank approximation of the kernel matrix, denoted as $K = \Phi\Phi^{\mathrm{T}}$ using random features [51]:

$$\begin{aligned} g_{N\times 1} &\sim \text{MVN}(\mathbf{0}_{N\times 1}, \boldsymbol{\Phi}\boldsymbol{\Phi}^T_{N\times N}) \\ \text{where} \quad \boldsymbol{\Phi}_{i,D_L\times 1} &= \sqrt{2/D_L} * \cos(\boldsymbol{W}_L h_i + \boldsymbol{b}_L) \end{aligned} \tag{7}$$

where $h_i = h(\boldsymbol{x}_i)$ represents the hidden representation in the hidden layer h_{L-1} with a dimension of D_{L-1}. Φ_i is the penultimate layer with dimension D_L, which consists of $\boldsymbol{W}_L$ and $\boldsymbol{b}_L$, shown in Eq. (7). The $\boldsymbol{W}_L$ is a fixed weight matrix with dimensions $D_L \times D_{L-1}$, and its entries are generated independently and identically randomly from a normal distribution with a mean of 0 and a standard deviation of 1, $\boldsymbol{W} \sim \mathcal{N}(0, 1)$. Additionally, the $\boldsymbol{b}_L$ is a fixed bias term with dimensions $D_{L\times 1}$, and its entries are independently sampled from a uniform distribution ranging between 0 and 2π, $\boldsymbol{b}_L \sim$ *Uniform(0,* 2π*)*.

Consequently, the RFF approximation to the GP prior in Eq. (7) for the k-th logit can be represented as a neural network layer. This layer has fixed hidden weights $\boldsymbol{W}_L$ and fixed biases $\boldsymbol{b}_L$, and the output weights denoted as β_k that can be learned:

$$\begin{aligned} g_k(h_i) &= \sqrt{2/D_L} * \cos(\boldsymbol{W}_L h_i + \boldsymbol{b}_L)^\top \beta_k \\ \text{with prior } \boldsymbol{\beta}_{k,D_L\times 1} &\sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}_{D_L\times D_L}) \end{aligned} \tag{8}$$

With the RFF approximation, conditional on the $h(\boldsymbol{x})$, $\boldsymbol{\beta} = \{\beta_k\}_{k=1}^{K}$ is the only trainable parameter in the model. Consequently, the RFF approximation in Eq. (8) transforms an infinite-dimensional GP into a finite Bayesian linear model. $\boldsymbol{\beta}$ together with the weights and bias in the $L-1$ hidden layers constitute the trainable parameters of the entire neural network. These parameters are trained to minimize the cross-entropy loss in the classification problem.

C. Uncertainty-Informed Risk Management

In real-world applications, DNN models are expected to accurately express the uncertainty associated with their predictions, especially when facing common open-world novelties, such as distribution shifts and OOD. Accurate and reliable predictive uncertainty estimations enable decision-makers to quantitatively know the potential risks involved in the decision-making under a certain scenario. Since we have limited control over what is fed into the deployed model, uncertainties such as distributional shifts arising from the constantly changing external environment pose significant risks to the model's predictions [56].

Performance of the deployed model might deteriorate over time and generate unreliable predictions due to variations in the underlying data distributions. Note that risk is reflected in the uncertainty regarding its associated consequences under a certain scenario [17, 26]. As shown in Fig. 3, we establish a unified uncertainty-informed risk assessment and management strategy to safeguard the DNN's applications in the presence of the two emerging uncertainties in the open environment - distribution shift and OOD. Specifically, for these three common data scenarios, we derive customized uncertainty thresholds (i.e. low and high uncertainty) to determine the degree of uncertainty in the model prediction for a specific sample. (1) The low uncertainty threshold U_l is selected using the Youden's Index, which minimizes the classification error on the validation set (shown as the upper-right subplot). In the subplot, the vertical dotted line marks the threshold separating low- and high-uncertainty regions for in-distribution data, whereas the dashed curve traces the corresponding Youden's Index values. (2) The high uncertainty threshold U_h is determined as the maximum predictive uncertainty on the normal validation data. (3) The medium uncertainty is an interval formed by the low uncertainty U_l and the high uncertainty U_h. Thus, we obtain three uncertainty intervals (i.e., (0, U_l], (U_l, U_h), [U_h, ∞]) to assess the predictive uncertainties estimated by the deployed model on unobserved data.

Furthermore, the three uncertainty intervals serve to inform what actions to take next (i.e. trust model or human intervention) considering decision-makers' review and judgment (e.g., value-based and policy-related considerations) in a certain scenario. For example, in an objection identification problem, if the predictive uncertainty is less than the U_l and its associated consequences are affordable, we can choose to trust the model prediction in order to reduce the workload. In contrast, for a cancer diagnosis system, if the predictive uncertainty falls within the interval (U_l, U_h) or even [U_h, ∞], such a case should be flagged and passed over to medical experts for further examination. This is because the consequences of a wrong cancer diagnosis are unacceptable.

Such a risk-informed strategy combines human knowledge and the DNN model to make better decisions. This contributes to building more trustworthy and safe DNN models. If decision-makers understand the consequences associated with the decisions, predictive uncertainty can be a valuable indicator to increase the decision-makers' awareness of the underlying risk. Consequently, the decision-makers can take necessary actions to alleviate and manage the risk. In addition, the estimated uncertainty also can signal when to update and refine the deployed model. For instance, if the predictions of the DNN model with high uncertainties appear many times, such a case could trigger the retraining of the deployed model. In contrast, the standard DNN models are unable to provide such valuable information to manage these emerging risks and trigger the model update. Section IV-E will provide a concrete example

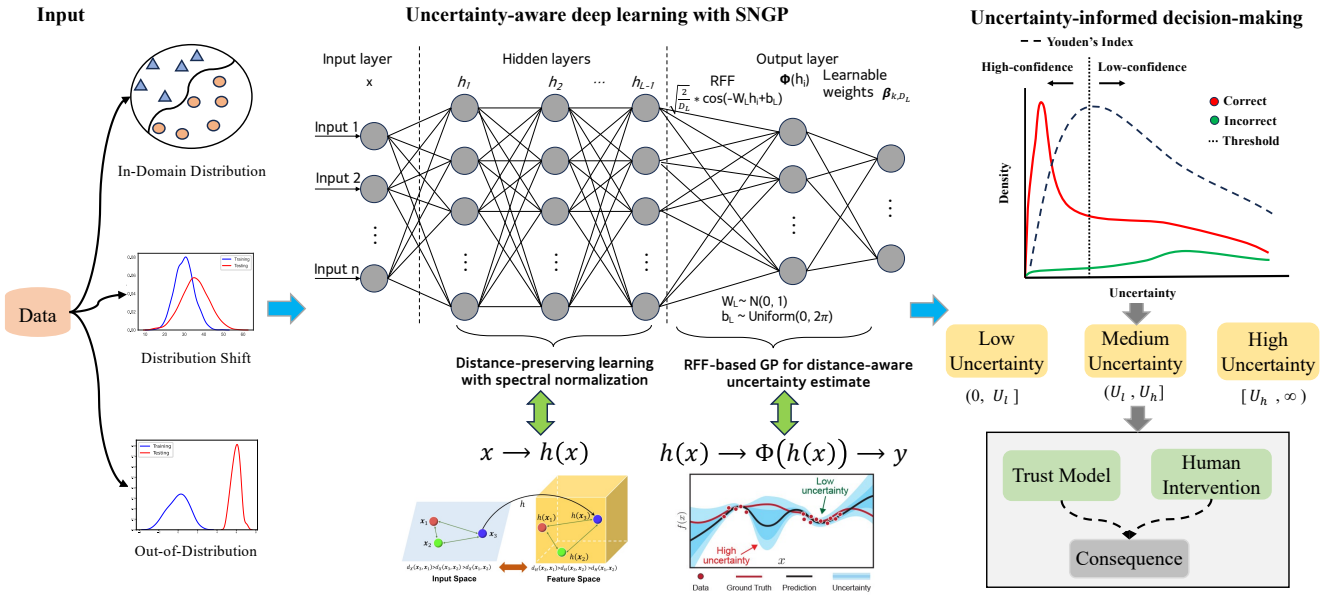


Fig. 3. Illustration of **uncertainty-informed risk assessment and management** of deep learning models under **distribution shift and Out-of-Distribution** [57]

to demonstrate the uncertainty-informed risk assessment and management strategy.

IV. Case Study

In this section, we perform two case studies to compare the prediction accuracy and the quality of the uncertainty estimated by the proposed method against MC dropout and deep ensembles. In addition, using the Diabetes example, we demonstrate how the estimated uncertainty facilitates risk assessment and management of deep learning models in practice.

A. Dataset Description

Diabetes is a chronic disease that affects about 38 million U.S. adults. In addition, approximately 98 million adults have prediabetes. Diabetes affects the human body in turning food into energy and also leads to other health issues, such as stroke, kidney failure, heart disease, etc, and even death. The economic cost of diabetes is substantial. For example, the cost of diagnosed diabetes in 2017 was around $327 million [58]. Early detection of diabetes has a significant impact by allowing for early medical intervention and potentially reducing the prevalence of the disease [31]. Prediabetes also has significant impacts both on health outcomes and quality of life, and early detection can help identify prediabetic individuals [58].

The applications of AI technologies in the treatment and study of diabetes have been extensively investigated in fundamental biomedical research, applied research, and clinical practice [59–61]. These technologies are helpful in the prediction of diabetes onset, management of risk factors, screening, automatic diet monitoring, etc. The predictive uncertainty can help decide whether to trust the model's prediction or take extra measures. Consequently, it can reduce and even avoid incorrect diagnoses. This will facilitate appropriately allocating health resources and reducing medical costs. More importantly, it also helps mitigate and avoid unnecessary physical and mental health pressure for a healthy person caused by a wrong diagnosis.

This paper considers a real-world binary classification task using diabetes data collected and published by the Centers for Disease Control and Prevention (CDC) [62]. Based on the data, Gardner et al. [31] extracted a set of features related to diabetes (e.g., general physical health, BMI/obesity, and demographic indicators) to develop a distribution shift benchmark termed as *TableShift*. To study the model's robustness to the distribution shift, they partition the data by race/ethnicity, where they utilize the "White non-Hispanic" individuals as the training data and all other race/ethnicity as the testing data. This is because all other race-ethnicity groups have a higher risk than "White non-Hispanic" individuals [62].

We fetch the data *diabetes* from the distribution shift benchmark. As Fig. 4 shows, we consider three different types of data for testing purpose, namely normal, shift, and OOD, to validate the performance of the developed model in UQ and classification accuracy. Apart from the OOD, all the datasets are drawn from the benchmark. We construct the OOD data based on the heart attack data by adding some trivial artificial features. More detailed information about the benchmark *TableShift* and data *diabetes* can be found in Ref. [31].

In addition to the diabetes dataset, we also evaluate the proposed method on the MNIST handwritten digit dataset [63] to assess its effectiveness when applied to convolutional neural networks (CNNs). We follow the method in Ref. [36] to create the shifted and OOD data by rotating the images and rolling the image pixel value in the normal test data.

B. Benchmark UQ Methods

1) **Monte Carlo dropout**: The MC dropout randomly drops a certain percentage of neurons during inferencing to generate a distribution of predictions. The posterior predictive

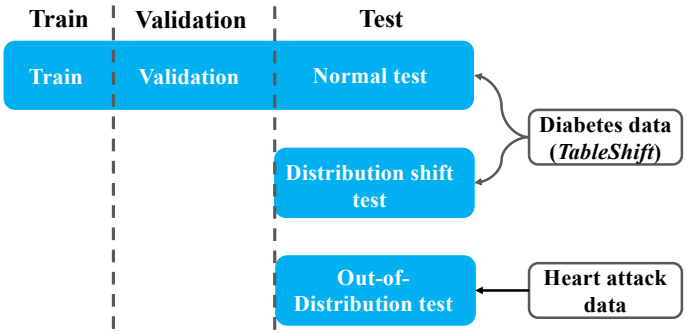


Fig. 4. **Experimental setup.** Models are trained on the data "Train" from the data *diabetes*. Predictions are made for three different test sets: (1) Normal test; (2) Distribution shift test; (3) Out-of-Distribution test.

distribution is approximated as follows [11]:

$$p(y^*|\boldsymbol{x}^*, D) = \int p(y^*|\boldsymbol{x}^*, \theta) \cdot p(\theta|D)\, d\theta \approx \frac{1}{N}\sum_{n=1}^{N} p(y^*|\boldsymbol{x}^*, \theta^{(n)}) \tag{9}$$

where D refers to the training data, $\boldsymbol{x}^*$ indicates the test sample, N means the number of forward passes with MC dropout, and $\theta^{(n)}$ indicates the i-th model's parameters. The posterior distribution is approximated by combining the outputs from N multiple forward passes. The variance of N predictions is used to model the predictive uncertainty.

2) **Deep ensembles**: In deep ensembles, multiple models with the same architecture are trained independently and then their predictions are combined together to estimate the predictive uncertainty[10]. Note that these models are trained with different configurations, such as parameter initialization and dataset splitting. The M predictions are combined to form a mixture distribution [11]:

$$p(y^*|\boldsymbol{x}^*) \approx \frac{1}{M}\sum_{m=1}^{M} p(y^*|\boldsymbol{x}^*, \theta_m) \tag{10}$$

where θ_m is the m-th independent model's parameters. The variance of these M predictions can be used to measure the uncertainty.

C. Performance Metrics

1) **Classification accuracy:** To examine the classification performance of the proposed method, we evaluate the prediction accuracy. This metric is used to benchmark the four considered methods, including the standard deep learning model, MC dropout, deep ensembles, and the proposed method.
2) **Correlation coefficient:** To assess the consistency and quality of the estimated uncertainty, we use the Pearson correlation coefficient to analyze the relationship between the quantified uncertainty and the hidden distance. The estimated uncertainty is expected to correlated with the degree of distribution shift. Specifically, as the distribution shift intensifies, the estimated uncertainty should also increase to reflect the growing distance between the in-distribution training data and the shifted data. Since the true uncertainty is unknown, the consistency of a model's estimated uncertainty becomes a key criterion for evaluating its quality. The predictive uncertainty should increase accordingly as the degree of distributional shift grows. Therefore, accurate and reliable uncertainty estimations are anticipated to align well with the distances between the training data and testing samples. Mathematically, we have the Pearson correlation coefficient r defined as [64]:

$$r = \frac{\sum_{i=1}^{n}(\sigma_i - \bar{\sigma})(d_i - \bar{d})}{\sqrt{\sum_{i=1}^{n}(\sigma_i - \bar{\sigma})^2 (d_i - \bar{d})^2}} \tag{11}$$

where σ_i and d_i indicate the predictive uncertainty (i.e., variance) and the corresponding distance indexed with i, $\bar{\sigma}$ and $\bar{d}$ are the mean of σ and d, respectively, and n is the total number of samples.
3) **Mean uncertainty:** The average prediction uncertainty across normal, shift, and OOD scenarios serves as an additional evaluation metric and helps address potential biases associated with the correlation coefficient. This metric offers an independent assessment of the consistency and reliability of our uncertainty estimates when the input data distribution changes. It offers an additional perspective on the quality of uncertainty estimation, complementing the insights obtained from the correlation with hidden distances.

D. Performance Comparison

We consider four methods - standard NN (no UQ capability), MC dropout, deep ensembles, and the proposed method - in the following computational study. Note that they have an identical residual neural network (ResNet) and CNN architectures for the two classification problems. Regarding the binary classification problem, the ResNet has 3 hidden layers, with each hidden layer consisting of 128 hidden units, a ReLU activation function, and a dropout rate of 0.1. The RestNet is trained to minimize the binary cross-entropy loss using the Adam optimizer with a learning rate of 10^{-4} and a batch size of 512 for 100 epochs. For the deep ensembles, 5 individual models are trained independently, and they produce 5 individual predictions for each test sample. In addition, we keep the same data setting from the distribution shift benchmark *TableShift* for the four considered methods [31]. Similarly, for the image data, the parameters on the CNN are kept the same across the four methods.

We compare the performance of the four considered methods from two aspects: classification accuracy and quality of estimated uncertainty. Note that variance is employed as an indicator of the predictive uncertainty of each method in the rest of the paper. In addition, the values of estimated uncertainty reported in Tables II and III are the average values under different shifts. Note the difference between the distributions of the shift test and the normal test is insignificant since only one feature *race* among 21 features is considered shifted [31]. Thus, we magnify the distances between the in-domain distribution and distributional shifts by adding Gaussian noises in the distribution shift and OOD. Such modification will highlight the distance-aware property of the proposed method.

Tables II and III summarize the performance of the four methods in terms of the predictive accuracy and the estimated

TABLE II
PERFORMANCE COMPARISON IN **CLASSIFICATION ACCURACY AND QUALITY OF ESTIMATED UNCERTAINTY** ON DATASET DIABETES

Scenario	**Classification accuracy**				**Estimated uncertainty**		
	Standard DNN ↑	MC dropout ↑	Deep ensembles ↑	Proposed method ↑	MC dropout	Deep ensembles	Proposed method
Non-shift	0.8735	0.8735	0.8733	0.8735	0.0006	0.0007	0.0085
Shift	0.8285	0.8291	0.8299	0.8291	0.0012	0.0046	0.0407
OOD	*	*	*	*	0.0003	0.0032	0.0598

TABLE III
PERFORMANCE COMPARISON IN **CLASSIFICATION ACCURACY AND QUALITY OF ESTIMATED UNCERTAINTY** ON MNIST

Scenario	**Classification accuracy**				**Estimated uncertainty**		
	Standard DNN↑	MC dropout↑	Deep ensembles↑	Proposed method↑	MC dropout	Deep ensembles	Proposed method
Non-shift	0.9918	0.9914	0.9939	0.9920	1.9757	2.0393	0.0048
Shift	0.5515	0.5898	0.6721	0.5818	1.5648	1.9661	0.0088
OOD	*	*	*	*	1.5153	1.7324	0.0102

TABLE IV
COMPARISON OF UNCERTAINTY ESTIMATION QUALITY USING NLL, BRIER, AND ECE ON THE MNIST DATASET

Scenario	Model	Metric		
		NLL ↓	Brier ↓	ECE ↓
Non-shift	Standard DNN	0.0304 ± 0.0043	0.0135 ± 0.0017	0.0047 ± 0.0008
	MC dropout	0.0284 ± 0.0039	0.0136 ± 0.0018	0.0025 ± 0.0006
	Deep ensembles	0.0169 ± 0.0007	0.0089 ± 0.0004	0.0020 ± 0.0004
	Proposed method	0.0252 ± 0.0023	0.0115 ± 0.0011	0.0021 ± 0.0006
	MC dropout (SNGP)	0.0272 ± 0.0021	0.0120 ± 0.0010	0.0046 ± 0.0007
	Deep ensembles (SNGP)	0.0170 ± 0.0003	0.0081 ± 0.0001	0.0034 ± 0.0001
Shift	Standard DNN	2.3450 ± 0.2808	0.6368 ± 0.0600	0.2703 ± 0.0292
	MC dropout	2.0305 ± 0.2352	0.6005 ± 0.0548	0.2141 ± 0.0281
	Deep ensembles	1.3598 ± 0.0495	0.4878 ± 0.0109	0.1097 ± 0.0050
	Proposed method	1.1605 ± 0.1004	0.4792 ± 0.0379	0.1618 ± 0.0188
	MC dropout (SNGP)	1.0699 ± 0.0865	0.4619 ± 0.0346	0.1078 ± 0.0184
	Deep ensembles (SNGP)	0.8375 ± 0.0192	0.3792 ± 0.0076	0.0443 ± 0.0078

TABLE V
RESULTS OF ONE-SIDED PAIRED T-TEST (SIGNIFICANCE THRESHOLD $p < 0.005$) FOR NLL, BRIER SCORE, AND ECE COMPARING THE PROPOSED METHOD AGAINST BASELINES UNDER NON-SHIFT AND SHIFT SCENARIOS

Scenario	Baseline	Comparison	P-value		
			NLL	Brier	ECE
Non-shift	Standard DNN	Proposed method	< 0.001	< 0.001	< 0.001
	MC dropout	Proposed method	0.0036	< 0.001	0.029
	MC dropout	MC dropout (SNGP)	0.126	< 0.001	1.000
	Deep ensembles	Deep ensembles (SNGP)	0.618	0.007	0.999
Shift	Standard DNN	Proposed method	< 0.001	< 0.001	< 0.001
	MC dropout	Proposed method	< 0.001	< 0.001	< 0.001
	MC dropout	MC dropout (SNGP)	< 0.001	< 0.001	< 0.001
	Deep ensembles	Deep ensembles (SNGP)	< 0.001	< 0.001	< 0.001

uncertainty. Clearly, the proposed method not only achieves comparable prediction accuracy, but also produces consistent uncertainty estimates. The predictive uncertainty grows as the degree of distributional shift grows. By contrast, the predictive uncertainties of the other two methods decrease on OOD data, indicating that we cannot rely on their uncertainty estimates to detect OOD data. Under the three considered scenarios, each method's estimated uncertainty varies. To further investigate this issue, we visualize the histograms of predictive variance estimated by the three methods in Fig. 5. As can be seen, the proposed method accurately distinguishes among the three data scenarios and effectively reflects these differences in its estimated uncertainties. Clearly, there is a distinct separation between the normal, distribution shift, and OOD data in the uncertainty estimated by the proposed method. However, MC dropout and deep ensembles fail to separate the three data scenarios as the estimated uncertainties are mixed up together.

In addition, we evaluate the quality of uncertainty estimates of these methods on both the non-shift and shifted MNIST datasets using three established metrics—Negative Log-Likelihood (NLL), Brier score, and Expected Calibration Error (ECE)—as summarized in Table IV. Specifically, we train 20 standard DNNs and 20 instances of the developed models, using identical dataset splits and initializing each model with the same seed for each independent run. For both the standard DNN and the proposed method, we report the mean and

standard deviation of the NLL, Brier score, and ECE values across the 20 independent runs. For MC dropout, each model randomly drops 10% of units during model inference, and the predictive mean is estimated by averaging over five stochastic dropout-enabled forward passes. We repeat this procedure across the twenty independently trained models to derive the mean and standard deviation for each method. For ensembles, we constructed five ensembles using bootstrap sampling, with each ensemble consisting of ten independently trained models. We report the mean and standard deviation of model performance across these five ensembles. We further evaluate the statistical significance of the performance differences using a paired t-test, as reported in Table V. It should be noted that $p < 0.001$ demonstrates that our proposed method achieves statistically significant improvements over the baseline. Based on the computational results presented in these two tables, we draw the following conclusions regarding predictive performance and the calibration quality of uncertainty estimation.

- **Proposed method vs. standard baselines**
  - **Non-shift setting:** The proposed method outperforms the standard DNN and MC dropout across all evaluation metrics, while it exhibits a slightly higher NLL and Brier score and comparable ECE relative to deep ensembles. For example, compared to the standard DNN, the proposed method achieves a lower NLL (0.0252 vs. 0.0304), Brier score (0.0115 vs. 0.0135), and ECE (0.0021 vs. 0.0047), with all improvements being statistically significant ($p < 0.001$). Compared to deep ensembles, however, it shows a slightly higher NLL (0.0252 vs. 0.0169) and Brier score (0.0115 vs. 0.0089), with a comparable ECE value (0.0021 vs. 0.0020).
  - **Shift setting:** The proposed method outperforms both the standard DNN and MC dropout across all metrics, with all improvements being statistically significant (p-value$<$0.001). Compared to deep ensembles, it achieves slightly lower Brier (0.4792 vs. 0.4878) and NLL (1.1605 vs. 1.3598), but higher ECE (0.1618 vs. 0.1097), reflecting a trade-off between accuracy and calibration.

Beyond the single-model variant, our proposed method can be seamlessly integrated with MC dropout and deep ensembles. As a result, we further compare the MC dropout and ensemble variants of SNGP against their respective counterparts.

- **MC dropout and ensembles of SNGP vs. their respective counterparts**
  - **Non-shift setting:** The SNGP-based MC dropout and deep ensemble achieve modest improvements in NLL and Brier scores over their baselines, but both incur higher ECE. For example, the SNGP-based MC dropout significantly reduces the Brier score (0.0120 vs. 0.0136, $p < 0.001$), and the SNGP-based deep ensemble also achieves a significantly lower Brier score (0.0081 vs. 0.0089, $p = 0.007$).
  - **Shift setting:** Both SNGP-based MC dropout and ensemble variants significantly outperform their counterparts across all evaluation metrics. Specifically, SNGP-based MC dropout improves in NLL (1.0699 vs. 2.0305), Brier (0.4619 vs. 0.6005), and ECE (0.1078 vs. 0.2141), all with p-value$<$0.001. The SNGP-based deep ensemble demonstrates the strongest overall performance, with substantial improvement in NLL (0.8375 vs. 1.3598), Brier (0.3792 vs. 0.4878), and ECE (0.0443 vs. 0.1097), all p-value$<$0.001.

In summary, these computational findings show that our proposed method achieves statistically significant improvements over both the standard DNN and MC dropout in the non-shift setting, with its advantages becoming even more pronounced under distribution shift. Importantly, the proposed method delivers performance comparable to deep ensembles while avoiding their heavy computational overhead. Requiring only a single forward pass, our method offers a scalable and computationally efficient solution for uncertainty estimation in practical settings. Moreover, the proposed method can be readily integrated with MC dropout and deep ensembles, resulting in superior performance compared to their standard counterparts. In addition, our approach reliably estimates the uncertainty for detecting OOD inputs - where ground-truth labels are unavailable and NLL, Brier, and ECE cannot be computed - highlighting its practical value for real-world deployment.

TABLE VI
PEARSON CORRELATION COEFFICIENT BETWEEN THE **HIDDEN DISTANCE AND ESTIMATED VARIANCE** ON DIABETES

Scenario	**Pearson correlation coefficient**			
	Standard DNN	MC dropout ↑	Deep ensembles↑	Proposed method↑
Non-shift	$*$	0.077	0.205	0.897
Shift	$*$	0.038	0.089	0.903
OOD	$*$	-0.198	-0.083	0.803
Overall	$*$	0.076	0.265	0.982

TABLE VII
PEARSON CORRELATION COEFFICIENT BETWEEN THE **HIDDEN DISTANCE AND ESTIMATED VARIANCE** ON DATASET MNSIT

Scenario	**Pearson correlation coefficient**			
	Standard DNN	MC dropout↑	Deep ensembles↑	Proposed method↑
Non-shift	$*$	0.641	0.612	0.802
Shift	$*$	0.614	0.463	0.868
OOD	$*$	0.603	0.470	0.908
Overall	$*$	0.654	0.536	0.850

TABLE VIII
TRAINING TIME ON TWO DIFFERENT DATASETS

Time(s)	**Training time on two different datasets**			
	Standard DNN↓	MC dropout↓	Deep ensembles↓	Proposed method ↓
Diabetes	20.35	20.35	93.24	27.65
MNIST	24.53	24.53	105.44	24.89

Table VI and VII present the Pearson correlation coefficients between hidden distances and estimated uncertainties across three data scenarios for each method. The hidden distance is the distance between the hidden representations of normal test data, shifted data, and OOD data, relative to their corresponding hidden representations of training data for each method. For example, to calculate the hidden distances between the training data and the shifted data for the proposed method: 1) we first derive the hidden representations of the training data using the proposed method; 2) we derive the hidden representations of the shifted data using the same method; 3) for each shifted data instance, we calculate its average Euclidean distance

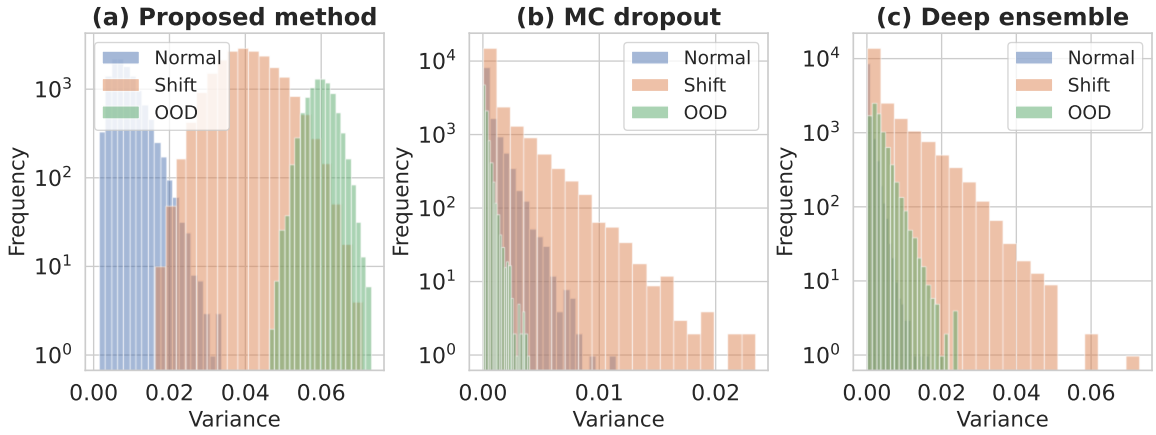


Fig. 5. Histogram of **estimated variances**

from the 10 nearest samples from the training data in the latent space. Note that we compute the input distance only for the tabular data to understand the relationship between the distributions of the input space distances and their associated hidden distances for each method. The input space distance, defined as the Euclidean distance between the training data and testing samples (e.g., normal, shifted, and OOD samples), is calculated without applying any feature extraction. Similar to the calculation of hidden distances, the final input distance for each testing sample is determined as the mean of the 10 samples with the nearest distances to the training data, as shown in Fig. 7(a).

Each column in Tables VI and VII displays the correlation between the hidden distance and the estimated uncertainty specific to that scenario. The last row (i.e., overall) reports the correlation across all three scenarios for each method. The proposed method achieves the highest correlation coefficient of 0.982 and 0.850, which suggests that the uncertainty estimated by the proposed method closely matches the hidden distance between the test samples and the training data. In fact, the uncertainty estimated by the proposed method aligns well with our expectations because it captures variations in the underlying data distribution. Specifically, as the data shift increases, the estimated uncertainty should rise accordingly to reflect the growing distance between the test samples and the training data. Notably, in both distribution shift and OOD settings, the proposed method consistently achieves significantly higher correlation coefficients compared to the other two methods across both datasets, which demonstrates the consistency and reliability in the UQ performance of the proposed method. These findings are further illustrated in Fig. 6.

As elaborated earlier in the Section III-B, distance is an important consideration in estimating the predictive uncertainty in the proposed method. On the one hand, the proposed method is able to preserve the distance of the input space in the hidden space by normalizing each hidden layer's weights with spectral normalization. Secondly, it leverages the relative distance in the hidden space to quantify the uncertainty by replacing the output layer with an RFF-approximated GP. Nevertheless, MC dropout and deep ensembles fail to distinguish the differences with respect to the distance between normal data and shift and OOD data. In this classification problem, MC dropout

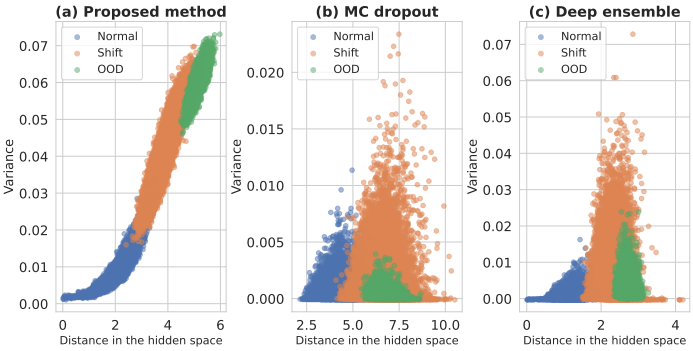


Fig. 6. Relationship between the **hidden distance** and **estimated variances**

and deep ensembles estimate the predictive uncertainty by measuring the distance between the testing sample and the classification boundary. Consequently, the correlation between the hidden distance and the estimated uncertainty of the proposed approach is much higher than the MC dropout and deep ensembles. The poor performance of MC dropout and the deep ensembles is attributed to their inability to preserve the relative distance among data points during data transformations.

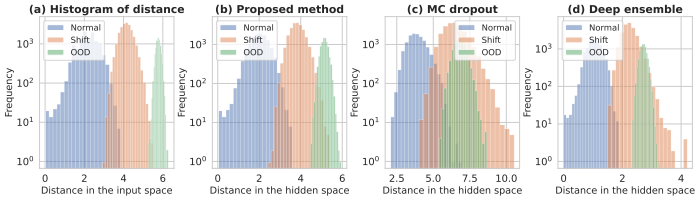


Fig. 7. Comparison of the **histogram of distance** (i.e. $\|h(\boldsymbol{x}) - h(\boldsymbol{x}')\|_H$)

To further verify this point, we compare the histograms of distance in the input space and the hidden space in Fig. 7. As it shows, the distance in the input space is well preserved in the hidden space after multiple transformations by the hidden layers in the proposed method. As the distributional shift grows, the associated distances increase. Since the proposed method is equipped with the distance-preserving property, the relative distance between the training data and the shift data is well retained in the hidden space, as shown in Fig. 7. However, the MC dropout and deep ensembles are not able to

maintain the distance in a principled and meaningful manner. By visualizing the distance in the input and hidden space, we further highlight the critical role of the distance-preserving property in uncertainty estimation.

Table VIII presents the training time of the four methods on tabular data (size: 87,230 × 142) and image data (size: 60,000 × 10). Evidently, the proposed method demonstrates comparable computational efficiency to the standard neural network. In contrast, deep ensembles suffer from the highest computational overhead as they need to train multiple individual models. The experiment was conducted using an NVIDIA GeForce RTX™ 4090. For implementation details and reproducibility materials, see our code repository at https://github.com/EricXue92/UI.

E. Uncertainty-Informed Risk Management

This section examines the application of UQ for uncertainty-informed risk assessment and management in diabetes prediction. Since the model development prior to deployment primarily relies on in-distribution data, we first use Youden's index, calculated from normal validation data [65], to determine a low uncertainty threshold U_l. Furthermore, the maximum predictive uncertainty, U_h, observed in the normal validation data, serves as a high uncertainty threshold to differentiate between in-distribution and OOD data.

Formally, Youden's index J is defined as follows:

$$J = \frac{\text{TP}}{\text{TP} + \text{FN}} + \frac{\text{TN}}{\text{TN} + \text{FP}} - 1 \tag{12}$$

where the first term, $\frac{\text{TP}}{\text{TP+FN}}$, represents true positive rate (sensitivity), and the second term, $\frac{\text{TN}}{\text{TN+FP}}$, represents true negative rate (specificity). Youden's index identifies the point on the receiver operating characteristic (ROC) curve that maximizes the trade-off between sensitivity and specificity. By balancing these two measures, Youden's index ensures that the chosen threshold minimizes the misclassification errors. The steps to calculate in distributionU_l associated with the maximum Youden's index are as follows:

1) **Generate the ROC curve**: The ROC curve is generated by plotting the predictive incorrectness (PI) against the predictive uncertainties (PU). Given the PI and the PU, the ROC curve can be computed as:

$$s_{\text{fpr}}, s_{\text{tpr}}, s_{\text{thresh}} = \text{roc\_curve}(\text{PI}, \text{PU}) \tag{13}$$

2) **Calculate Youden's index (J)**: The index is computed as the difference between the true positive rate (s_{tpr}) and the false positive rate (s_{fpr}):

$$J_i = s_{\text{tpr},i} - s_{\text{fpr},i} \tag{14}$$

3) **Determine the maximum Youden's index and its corresponding uncertainty threshold**: The optimal threshold is identified by finding the point where J is maximized. The uncertainty threshold U_l is as follows,

$$U_l = s_{\text{thresh}}\left(\arg\max_i J_i\right) \tag{15}$$

where $\arg\max_i$ identifies the index i where the difference $s_{\text{tpr},i} - s_{\text{fpr},i}$ is maximized, and $s_{\text{thresh}}(i)$ represents the threshold uncertainty corresponding to this optimal index.

In real-world settings, patient demographics can evolve — across time or between hospitals — leading to shifts in the underlying diabetes data distribution. Such shifts introduce varying levels of uncertainty in DNN predictions. By quantifying these levels, decision-makers can judge when to trust or question the model's output. As Fig. 8 shows, the vertical red dotted line indicates the uncertainty threshold U_l (i.e. 0.010) to distinguish the low and high uncertainty over the normal validation data. The black line represents the U_h (i.e. 0.034) to differentiate the OOD and in-distribution input. These two thresholds U_l and U_h enable decision-makers to gain a quantitative understanding of the uncertainty level and its potential corresponding uncertainty scenario. Usually, the risks caused by the OOD data associated with the uncertainty greater than the U_h are unacceptable as the input is beyond the scope of the trained model. When the uncertainty falls within the range $[U_l, U_h]$, the associated risk can be determined according to the specific situation, e.g., risk tolerance level, potential consequences, risk policy, etc. Generally, when the uncertainty is greater than the prescribed threshold U_l, users should be cautious and examine the particular DNN prediction carefully. By contrast, if the predictive uncertainty is less than the threshold U_l, we could choose to trust the model's prediction to reduce the workload.

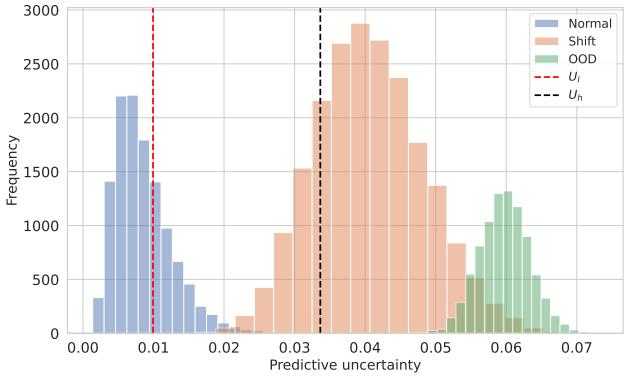


Fig. 8. Determination of **uncertainty thresholds**

V. Conclusion and Future Work

This paper leverages a distance-aware UQ method to estimate predictive uncertainty for DNNs under three scenarios: no distribution shift, distribution shift, and OOD inputs. We empirically evaluate classification accuracy and the quality of the uncertainty estimates against two widely used UQ baselines. Unlike these baselines, the proposed approach preserves pairwise distances throughout the feature transformations processes and uses an RFF-based GP to produce distance-sensitive uncertainty estimates. Consequently, it matches the predictive performance of standard NNs, MC dropout, and deep ensembles, while delivering more consistent and reliable uncertainty assessments. The observed correlations among input-space distances, hidden-space distances, and predictive variances further validate the importance of distance preservation.

We also demonstrated the value of the unified uncertainty-informed risk assessment and management strategy through a diabetes prediction case study. Beyond healthcare, the proposed UQ framework has strong potential in other high-stakes

applications. For example, in autonomous driving, it can detect novel or anomalous road conditions not present during training; in power systems, it can provide reliable uncertainty estimates for renewable energy generation to support the secure integration of renewables into the grid. Future work may explore these application domains more extensively. Additionally, a deeper examination of the approximation error introduced by RFF could further improve the fidelity of the uncertainty estimates [66].

ACKNOWLEDGMENTS

The work described in this paper is supported by a grant from the Research Grants Council of the Hong Kong Special Administrative Region, China (Project No. PolyU 25206422), the National Natural Science Foundation of China (Grant Nos. 62406269, 72271025), and the Research Committee of The Hong Kong Polytechnic University (Project code: RKB0, G-UARJ).

REFERENCES

[1] D. Feng, L. Rosenbaum, and K. Dietmayer, "Towards safe autonomous driving: Capture uncertainty in the deep neural network for lidar 3D vehicle detection," in *2018 21st International Conference on Intelligent Transportation Systems (ITSC)*. IEEE, 2018, pp. 3266–3273.

[2] J. Zheng, Y. Tang, A. Huang, and D. Wu, "Hierarchical multivariate representation learning for face sketch recognition," *IEEE Transactions on Emerging Topics in Computational Intelligence*, 2024.

[3] P. Suta, X. Lan, B. Wu, P. Mongkolnam, and J. H. Chan, "An overview of machine learning in chatbots," *International Journal of Mechanical Engineering and Robotics Research*, vol. 9, no. 4, pp. 502–510, 2020.

[4] Q. Chen, W. Wang, F. Wu, S. De, R. Wang, B. Zhang, and X. Huang, "A survey on an emerging area: Deep learning for smart city data," *IEEE Transactions on Emerging Topics in Computational Intelligence*, vol. 3, no. 5, pp. 392–410, 2019.

[5] X. Zhang, F. T. Chan, C. Yan, and I. Bose, "Towards risk-aware artificial intelligence and machine learning systems: An overview," *Decision Support Systems*, vol. 159, p. 113800, 2022.

[6] V. Nemani, L. Biggio, X. Huan, Z. Hu, O. Fink, A. Tran, Y. Wang, X. Zhang, and C. Hu, "Uncertainty quantification in machine learning for engineering design and health prognostics: A tutorial," *Mechanical Systems and Signal Processing*, vol. 205, p. 110796, 2023.

[7] C. Macrae, "Learning from the failure of autonomous and intelligent systems: Accidents, safety, and sociotechnical sources of risk," *Risk Analysis*, vol. 42, no. 9, pp. 1999–2025, 2022.

[8] X. Zhang, T. Wang, L. Ma, and S. Mahadevan, "Reliability engineering, risk management, and trustworthiness assurance for ai systems," *Journal of Reliability Science and Engineering*, vol. 1, no. 2, p. 022001, 2025.

[9] J. Gawlikowski, C. R. N. Tassi, M. Ali, J. Lee, M. Humt, J. Feng, A. Kruspe, R. Triebel, P. Jung, R. Roscher *et al.*, "A survey of uncertainty in deep neural networks," *Artificial Intelligence Review*, pp. 1–77, 2023.

[10] B. Lakshminarayanan, A. Pritzel, and C. Blundell, "Simple and scalable predictive uncertainty estimation using deep ensembles," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[11] A. Thuy and D. F. Benoit, "Explainability through uncertainty: Trustworthy decision-making with neural networks," *European Journal of Operational Research*, 2023.

[12] K. Lakara, A. Bhandari, P. Seth, and U. Verma, "Evaluating predictive uncertainty and robustness to distributional shift using real world data," *arXiv preprint arXiv:2111.04665*, 2021.

[13] E. Begoli, T. Bhattacharya, and D. Kusnezov, "The need for uncertainty quantification in machine-assisted medical decision making," *Nature Machine Intelligence*, vol. 1, no. 1, pp. 20–23, 2019.

[14] A. Malinin and M. Gales, "Predictive uncertainty estimation via prior networks," *Advances in Neural Information Processing Systems*, vol. 31, 2018.

[15] M. Chua, D. Kim, J. Choi, N. G. Lee, V. Deshpande, J. Schwab, M. H. Lev, R. G. Gonzalez, M. S. Gee, and S. Do, "Tackling prediction uncertainty in machine learning for healthcare," *Nature Biomedical Engineering*, vol. 7, no. 6, pp. 711–718, 2023.

[16] A. Bostrom, J. L. Demuth, C. D. Wirz, M. G. Cains, A. Schumacher, D. Madlambayan, A. S. Bansal, A. Bearth, R. Chase, K. M. Crosman *et al.*, "Trust and trustworthy artificial intelligence: A research agenda for ai in the environmental sciences," *Risk Analysis*, 2023.

[17] T. Aven, "Risk assessment and risk management: Review of recent advances on their foundation," *European Journal of Operational Research*, vol. 253, no. 1, pp. 1–13, 2016.

[18] ——, "The risk concept—historical and recent development trends," *Reliability Engineering & System Safety*, vol. 99, pp. 33–44, 2012.

[19] T. Aven, Y. Ben-Haim, H. Boje Andersen, T. Cox, E. L. Droguett, M. Greenberg, S. Guikema, W. Kröger, O. Renn, K. M. Thompson *et al.*, "Society for risk analysis glossary," in *Society for Risk Analysis*. Society for Risk Analysis, 2018, pp. 3–9.

[20] M. E. Paté-Cornell, "Uncertainties in risk analysis: Six levels of treatment," *Reliability Engineering & System Safety*, vol. 54, no. 2-3, pp. 95–111, 1996.

[21] R. L. Winkler, "Uncertainty in probabilistic risk assessment," *Reliability Engineering & System Safety*, vol. 54, no. 2-3, pp. 127–132, 1996.

[22] S. Kaplan and B. J. Garrick, "On the quantitative definition of risk," *Risk Analysis*, vol. 1, no. 1, pp. 11–27, 1981.

[23] J. C. Helton and F. Davis, "Illustration of sampling-based methods for uncertainty and sensitivity analysis," *Risk Analysis*, vol. 22, no. 3, pp. 591–622, 2002.

[24] R. T. Clemen and R. L. Winkler, "Combining probability distributions from experts in risk analysis," *Risk Analysis*, vol. 19, pp. 187–203, 1999.

[25] R. Flage, T. Aven, E. Zio, and P. Baraldi, "Concerns, challenges, and directions of development for the issue of representing uncertainty in risk assessment," *Risk Analysis*, vol. 34, no. 7, pp. 1196–1207, 2014.

[26] A. Terje, "On how to define, understand and describe risk," *Reliability Engineering & System Safety*, vol. 95, no. 6, pp. 623–631, 2010.

[27] S. Samson, J. A. Reneke, and M. M. Wiecek, "A review of different perspectives on uncertainty and risk and an alternative modeling paradigm," *Reliability Engineering & System Safety*, vol. 94, no. 2, pp. 558–567, 2009.

[28] Y. Zhang, P. Tiňo, A. Leonardis, and K. Tang, "A survey on neural network interpretability," *IEEE Transactions on Emerging Topics in Computational Intelligence*, vol. 5, no. 5, pp. 726–742, 2021.

[29] S. Gui, X. Li, L. Wang, and S. Ji, "Good: A graph out-of-distribution benchmark," *Advances in Neural Information Processing Systems*, vol. 35, pp. 2059–2073, 2022.

[30] P. W. Koh, S. Sagawa, H. Marklund, S. M. Xie, M. Zhang, A. Balsubramani, W. Hu, M. Yasunaga, R. L. Phillips, I. Gao *et al.*, "Wilds: A benchmark of in-the-wild distribution shifts," in *International Conference on Machine Learning*. PMLR, 2021, pp. 5637–5664.

[31] J. Gardner, Z. Popovic, and L. Schmidt, "Benchmarking distribution shift in tabular data with tableshift," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[32] A. Malinin, N. Band, G. Chesnokov, Y. Gal, M. J. Gales, A. Noskov, A. Ploskonosov, L. Prokhorenkova, I. Provilkov, V. Raina *et al.*, "Shifts: A dataset of real distributional shift across multiple large-scale tasks," *arXiv preprint arXiv:2107.07455*, 2021.

[33] A. Malinin, A. Athanasopoulos, M. Barakovic, M. B. Cuadra, M. J. Gales, C. Granziera, M. Graziani, N. Kartashev, K. Kyriakopoulos, P.-J. Lu *et al.*, "Shifts 2.0: Extending the dataset of real distributional shifts," *arXiv preprint arXiv:2206.15407*, 2022.

[34] L. H. Nazer, R. Zatarah, S. Waldrip, J. X. C. Ke, M. Moukheiber, A. K. Khanna, R. S. Hicklen, L. Moukheiber, D. Moukheiber, H. Ma *et al.*, "Bias in artificial intelligence algorithms and recommendations for mitigation," *PLOS Digital Health*, vol. 2, no. 6, p. e0000278, 2023.

[35] W. Liang, G. A. Tadesse, D. Ho, L. Fei-Fei, M. Zaharia, C. Zhang, and J. Zou, "Advances, challenges and opportunities in creating data for trustworthy AI," *Nature Machine Intelligence*, vol. 4, no. 8, pp. 669–677, 2022.

[36] Y. Ovadia, E. Fertig, J. Ren, Z. Nado, D. Sculley, S. Nowozin, J. Dillon, B. Lakshminarayanan, and J. Snoek, "Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[37] M. Abdar, F. Pourpanah, S. Hussain, D. Rezazadegan, L. Liu, M. Ghavamzadeh, P. Fieguth, X. Cao, A. Khosravi, U. R. Acharya *et al.*, "A review of uncertainty quantification in deep learning: Techniques, applications and challenges," *Information Fusion*, vol. 76, pp. 243–297, 2021.

[38] X. Zhang and S. Mahadevan, "Bayesian neural networks for flight trajectory prediction and safety assessment," *Decision Support Systems*, vol. 131, p. 113246, 2020.

[39] A. Graves, "Practical variational inference for neural networks," *Advances in Neural Information Processing Systems*, vol. 24, 2011.

[40] M. David, "Bayesian methods for adaptive models," *Doctoral Thesis, California Institute of Technology*, 1992.

[41] Y. Gal and Z. Ghahramani, "Dropout as a bayesian approximation: Representing model uncertainty in deep learning," in *International Conference on Machine Learning*. PMLR, 2016, pp. 1050–1059.

[42] J. Liu, Z. Lin, S. Padhy, D. Tran, T. Bedrax Weiss, and B. Lakshminarayanan, "Simple and principled uncertainty estimation with deterministic deep learning via distance awareness," *Advances in Neural Information Processing Systems*, vol. 33, pp. 7498–7512, 2020.

[43] J. Z. Liu, S. Padhy, J. Ren, Z. Lin, Y. Wen, G. Jerfel, Z. Nado, J. Snoek, D. Tran, and B. Lakshminarayanan, "A simple approach to improve single-model deep uncertainty via distance-awareness." *J. Mach. Learn. Res.*, vol. 24, pp. 42–1, 2023.

[44] W. Yi, W. K. Chan, H. H. Lee, S. T. Boles, and X. Zhang, "An uncertainty-aware deep learning model for reliable detection of steel wire rope defects," *IEEE Transactions on Reliability*, pp. 1–15, 2023.

[45] A. NS, M. S, and M. J., "Modeling uncertainty in risk assessment: An integrated approach with fuzzy set theory and monte carlo simulation," *Accident Analysis & Preventionh*, vol. 55, no. 13, pp. 242–255, 2013.

[46] X. Zhou, H. Liu, F. Pourpanah, T. Zeng, and X. Wang, "A survey on epistemic (model) uncertainty in supervised learning: Recent advances and applications," *Neurocomputing*, vol. 489, pp. 449–465, 2022.

[47] F. Coolen and L. Utkin, "Imprecise reliability: A concise overview," in *Proceedings of the European Safety and Reliability Conference 2007, ESREL 2007-Risk, Reliability and Societal Safety*. ESREL 2007, 2007, pp. 1959–1966.

[48] W. He, Z. Jiang, T. Xiao, Z. Xu, and Y. Li, "A survey on uncertainty quantification methods for deep learning," *ACM Computing Surveys*, 2025.

[49] F. Bayram and B. S. Ahmed, "A domain-region based evaluation of ml performance robustness to covariate shift," *Neural Computing and Applications*, pp. 1–23, 2023.

[50] J. Van Amersfoort, L. Smith, A. Jesson, O. Key, and Y. Gal, "On feature collapse and deep kernel learning for single forward pass uncertainty," *arXiv preprint arXiv:2102.11409*, 2021.

[51] A. Rahimi and B. Recht, "Random features for large-scale kernel machines," *Advances in Neural Information Processing Systems*, vol. 20, 2007.

[52] J. Van Amersfoort, L. Smith, Y. W. Teh, and Y. Gal, "Uncertainty estimation using a single deep deterministic neural network," in *International Conference on Machine Learning*. PMLR, 2020, pp. 9690–9700.

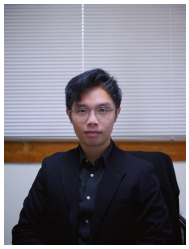
[53] M. O'Searcoid, *Metric spaces*. Springer Science & Business Media, 2006.

[54] H. Gouk, E. Frank, B. Pfahringer, and M. J. Cree, "Regularisation of neural networks by enforcing lipschitz continuity," *Machine Learning*, vol. 110, pp. 393–416, 2021.

[55] T. Miyato, T. Kataoka, M. Koyama, and Y. Yoshida, "Spectral normalization for generative adversarial networks," *arXiv preprint arXiv:1802.05957*, 2018.

[56] L. A. Cox Jr, "Answerable and unanswerable questions in risk analysis with open-world novelty," *Risk Analysis*, vol. 40, no. S1, pp. 2144–2177, 2020.
[57] P. Bandi, O. Geessink, Q. Manson, M. Van Dijk, M. Balkenhol, M. Hermsen, B. E. Bejnordi, B. Lee, K. Paeng, A. Zhong *et al.*, "From detection of individual metastases to classification of lymph node status at the patient level: the camelyon17 challenge," *IEEE Transactions on Medical Imaging*, vol. 38, no. 2, pp. 550–560, 2018.
[58] A. D. Association, "Economic costs of diabetes in the us in 2017," *Diabetes Care*, vol. 41, no. 5, pp. 917–928, 2018.
[59] Z. Guan, H. Li, R. Liu, C. Cai, Y. Liu, J. Li, X. Wang, S. Huang, L. Wu, D. Liu *et al.*, "Artificial intelligence in diabetes management: advancements, opportunities, and challenges," *Cell Reports Medicine*, 2023.
[60] I. Kavakiotis, O. Tsave, A. Salifoglou, N. Maglaveras, I. Vlahavas, and I. Chouvarda, "Machine learning and data mining methods in diabetes research," *Computational and Structural Biotechnology Journal*, vol. 15, pp. 104–116, 2017.
[61] S. Siuly, Ö. F. Alçin, H. Wang, Y. Li, and P. Wen, "Exploring rhythms and channels-based eeg biomarkers for early detection of alzheimer's disease," *IEEE Transactions on Emerging Topics in Computational Intelligence*, 2024.
[62] Centers for Disease Control and Prevention, "National diabetes statistics report," https://www.cdc.gov/diabetes/data/statistics-report/index.html, 2022, accessed: Jan. 05, 2023.
[63] L. Deng, "The mnist database of handwritten digit images for machine learning research [best of the web]," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 141–142, 2012.
[64] I. Cohen, Y. Huang, J. Chen, J. Benesty, J. Benesty, J. Chen, Y. Huang, and I. Cohen, "Pearson correlation coefficient," *Noise Reduction in Speech Processing*, pp. 1–4, 2009.
[65] J. M. Dolezal, A. Srisuwananukorn, D. Karpeyev, S. Ramesh, S. Kochanny, B. Cody, A. S. Mansfield, S. Rakshit, R. Bansal, M. C. Bois *et al.*, "Uncertainty-informed deep learning models enable high-confidence predictions for digital histopathology," *Nature Communications*, vol. 13, no. 1, p. 6572, 2022.
[66] J. Yao, N. B. Erichson, and M. E. Lopes, "Error estimation for random fourier features," in *International Conference on Artificial Intelligence and Statistics*. PMLR, 2023, pp. 2348–2364.

Long Xue is a Ph.D. candidate in the Department of Industrial and Systems Engineering at The Hong Kong Polytechnic University. He received his M.S. in Control Science and Engineering from the School of Reliability and Systems Engineering, Beihang University (Beijing, China). His research focuses on uncertainty quantification and trustworthy AI, including conformal prediction, out-of-distribution detection, and robustness under distribution shift. He is particularly interested in developing methods that provide reliable uncertainty estimates and finite-sample coverage guarantees for deep learning models in open-world settings. In 2025, he was a visiting Ph.D. researcher at the University of Warwick (UK), where he explored robust conformal prediction under adversarial attacks.



Sai-Ho Chung (Nick) Ph.D., is an Associate Head and Associate Professor of Industrial and Systems Engineering at the Hong Kong Polytechnic University. He obtained his Ph.D. degree from the University of Hong Kong. His research interests are logistics, supply chain management, supply chain finance, production scheduling, distribution networks, machine learning, container port terminals, aviation, etc. He has published over 150 SCI journal papers. His publications appear in POM, TR Part B/C/E, DS, RA, EJOR, IEEE (SMC, TIE, EM, SJ), DSS, IJPE, IJPR, COR, etc. Among which, the paper published in RA was selected as the top paper, the media of the month by Society for Risk Analysis, and the top 0.1 paper in Almetric. In Google Scholar, he currently has over 10,000 citations with H-index 52. He has been ranked in the World's Top 2% Scientists on Stanford University List. Moreover, he has obtained 2 patents and numerous research funding with a total amount of over HK$15M. He serves as a Department Editor for IEEE TEM, and Associate Editor for IEEE TSMC, and TRE. He edited several special issues in SCI journals.

Lechang Yang received a bachelor's degree in aircraft design and a doctoral degree in systems engineering from Beihang University, Beijing, China. He is now a full professor in the School of Mechanical Engineering, University of Science and Technology Beijing. His research interests include reliability engineering, uncertainty quantification, and artificial intelligence algorithm.

Xiao-Lin Wang is an associate professor in the Business School at Sichuan University, Chengdu, China. Prior to that, he was a research assistant professor in the Department of Logistics and Maritime Studies at The Hong Kong Polytechnic University, Hong Kong SAR. He received his PhD in industrial engineering from City University of Hong Kong in 2020, and his BS and MS degrees in industrial engineering from Southeast University, Nanjing, China, in 2013 and 2016, respectively. His research interest lies in applying data analytics, stochastic modeling, and optimization techniques to solve maintenance optimization, warranty analytics, and operations management problems. His research outcomes have appeared in IEEE Transactions on Reliability, IISE Transactions, INFORMS Journal on Computing, European Journal of Operational Research, Manufacturing & Service Operations Management, among others.

Xiaoge Zhang is an Assistant Professor in the Department of Industrial and Systems Engineering (ISE) at The Hong Kong Polytechnic University. He received his PhD in Systems Engineering and Operations Research from Vanderbilt University, Nashville, Tennessee, United States in 2019. He has won multiple awards, including Peter G. Hoadley Best Paper Award, Chinese Government Award for Outstanding Self-Financed Students Studying Abroad, Bravo Zulu Award, Pao Chung Chen Fellowship, among others. He has published more than 90 research papers in leading academic journals, such as Nature Communications, INFORMS Journal on Computing, IEEE Transactions on Automation Science and Engineering, IEEE Transactions on Intelligent Transportation Systems, Reliability Engineering & Systems Safety, Risk Analysis, IEEE Transactions on Industrial Informatics, IEEE Transactions on Reliability, Decision Support Systems, IEEE Transactions on Cybernetics, and Annals of Operations Research, among others. His research has gathered widespread attention from the academic community (4700+ citation, h-index 35 according to Google Scholar). His research interests center on reliable AI, machine learning, reliability modeling of autonomous systems, and uncertainty quantification. He is a senior member of IEEE and a member of INFORMS and IISE.