Towards Trustworthy Vital Sign Forecasting: Leveraging Uncertainty for Prediction Intervals

1st Li Rong Wang
College of Computing and Data Science
Nanyang Technological University
*& A*STAR Centre for Frontier AI Research*
Singapore
0000-0003-3546-9354

2nd Thomas C. Henderson
School of Computing
University of Utah
Salt Lake City, United States
0000-0002-0792-3882

3rd Yew Soon Ong
College of Computing and Data Science
Nanyang Technological University
*& A*STAR Centre for Frontier AI Research*
Singapore
0000-0002-4480-169X

4th Yih Yng Ng
Swee Hock School of Public Health
National University of Singapore
Singapore
0000-0003-4598-1829

5th Xiuyi Fan
Lee Kong Chian School of Medicine
& College of Computing and Data Science
Nanyang Technological University
Singapore
0000-0003-1223-9986

***Abstract*—Vital signs, such as heart rate and blood pressure, are critical indicators of patient health and are widely used in clinical monitoring and decision-making. While deep learning models have shown promise in forecasting these signals, their deployment in healthcare remains limited in part because clinicians must be able to trust and interpret model outputs. Without reliable uncertainty quantification – particularly calibrated prediction intervals (PIs) – it is unclear whether a forecasted abnormality constitutes a meaningful warning or merely reflects model noise, hindering clinical decision-making. To address this, we present two methods for deriving PIs from the Reconstruction Uncertainty Estimate (RUE), an uncertainty measure well-suited to vital-sign forecasting due to its sensitivity to data shifts and support for label-free calibration. Our parametric approach assumes that prediction errors and uncertainty estimates follow a Gaussian copula distribution, enabling closed-form PI computation. Our non-parametric approach, based on k-nearest neighbours (KNN), empirically estimates the conditional error distribution using similar validation instances. We evaluate these methods on two large public datasets with minute- and hour-level sampling, representing high- and low-frequency health signals. Experiments demonstrate that the Gaussian copula method consistently outperforms conformal prediction baselines on low-frequency data, while the KNN approach performs best on high-frequency data. These results underscore the clinical promise of RUE-derived PIs for delivering interpretable, uncertainty-aware vital sign forecasts.**

***Index Terms*—Prediction interval generation, vital sign forecasting, Uncertainty estimation, Uncertainty quantification**

I. Introduction

Vital signs, such as heart rate, respiratory rate, and blood pressure, are fundamental clinical measurements that provide critical information about patients' health and body functions [1], [2]. They are sensitive indicators of a patient's condition and are often the first signals of physiological distress [3], [4]. Their abnormalities can precede visible symptoms, making them essential for early intervention [5], [6], especially in acute care, emergency medicine, intensive care, and hospital at home settings.

Recently, machine learning methods have been explored to *predict* vital signs [7]–[9]. These approaches represent a shift from treating vital signs solely as inputs for risk score calculations to directly forecasting their future trajectories using patients' electronic medical records. Despite promising results, almost all studies report only point-error metrics (e.g., MAE, RMSE), leaving clinicians unaware of the reliability of each forecast. Without calibrated *Prediction Intervals (PIs)*, it is impossible to determine whether a projected hypertensive episode constitutes a high-confidence warning or merely reflects noise in the model. Given the high-stakes nature of clinical decision-making, the absence of uncertainty quantification substantially limits the practical utility of these models.

To address this gap, we propose two PI estimation method based on uncertainty quantification, enabling vital sign prediction models to be better equipped for clinical deployment. Unlike traditional uncertainty quantification approaches such as Bayesian Neural Networks (BNNs) [10], Monte Carlo dropout (MCD) [11], and deep ensembles [12], which typically produce a single scalar "uncertainty score" for each prediction, our methods generate PIs that explicitly represent a range of plausible outcomes. The width of each interval intuitively reflects the model's confidence in its forecast. For example, rather than outputting an abstract "uncertainty value of 5" for a heart rate prediction, our method would output a concrete interval such as 160 to 164 beats per minute.

For a patient $\mathbf{x}$, let a vital sign forecasting model produce a point prediction $\hat{\mathbf{y}}$, and let an *uncertainty estimation* model $u(\mathbf{x})$ return a confidence measure. Given a target error tolerance level $\alpha \in (0, 1)$, (typically $\alpha = 0.05$), we define a

This research was supported by the Ministry of Education, Singapore (Grant ID: RS15/23) and the College of Computing and Data Science, Nanyang Technological University.

prediction interval by estimating the smallest ϵ such that

$$\Pr\left(|\mathbf{y} - \hat{\mathbf{y}}| \leq \epsilon \mid u(\mathbf{x})\right) = 1 - \alpha, \tag{1}$$

where $\mathbf{y}$ is the true value of the vital sign in the future. The resulting interval $[\hat{\mathbf{y}} - \epsilon, \hat{\mathbf{y}} + \epsilon]$ achieves the prescribed coverage level while adapting its width to the model's uncertainty $u(\mathbf{x})$.

A widely used method for deriving PIs is Conformal Prediction (CP). Conventional CP [13] sets the PI width ϵ to the appropriate quantile of absolute prediction errors (known as the non-conformity score) measured on a calibration set. Normalized CP [14]–[17] extends this approach by normalizing the non-conformity score with a scalar uncertainty estimate, producing an uncertainty-conditioned PI. Building on this strong foundation, we enhance normalized CP's ability to model Equation 1 with two key improvements:

1) **Richer uncertainty-width mapping**. We allow more expressive nonlinear functions – such as Gaussian copulas and K-nearest neighbours (KNN) —- beyond simple scalar multiplication to capture complex relationships between uncertainty estimates $u(\mathbf{x})$ and interval half-width ϵ.
2) **Multi-dimensional conditional.** We generalize CP to handle vector-valued uncertainty estimates $u(\mathbf{x}) \in \mathbb{R}^K, K > 1$, enabling tighter, context-aware intervals.

To achieve this, we employed the recently introduced multi-dimensional uncertainty measure, Reconstruction Uncertainty Estimate (RUE) [18], [19] for computing $u(\mathbf{x})$. RUE has seen its success in time-series applications [20], [21] and is designed for quantifying uncertainty from data shifts. RUE is particularly well suited to vital-sign forecasting because it supports label-free calibration, allowing continuous updates from unlabelled telemetry streams. Its reconstruction-error mechanism also remains sensitive to the covariate shifts common in bedside monitoring, such as sensor drift, changes in patient mix, and irregular sampling. These properties enable us to build reliable, context-aware prediction intervals without repeatedly retraining the base forecaster.

Leveraging on the multiple-dimensional outputs from RUE, we developed two strategies to derive PIs from RUE. The parametric approach, based on a Gaussian copula, assumes that the uncertainty scores and prediction errors follow a joint distribution, enabling closed-form computation of the conditional distribution in Equation (1). The non-parametric approach, a k-nearest-neighbours (KNN) estimator, constructs this conditional distribution empirically by using prediction errors from validation points with similar reconstruction errors.

In vital sign prediction, a wide range of health signals are used, with sampling rates spanning from high frequencies (e.g., ECG at 1024 Hz and accelerometers at 50 Hz [22], [23]) to low frequencies measured in days (e.g., steps, heart rate, sleep status, and blood oxygen saturation [24]). We evaluate our PI methods on two large public vital signs datasets sampled at minute and hour intervals to test their effectiveness across different frequencies. Our experiments demonstrate that the Gaussian-copula approach consistently surpasses CP baselines across several evaluation metrics—particularly on low-frequency data—whereas the KNN method achieves superior performance on high-frequency data.

II. Related Work

a) Vital Sign Forecasting: Recent work on vital sign forecasting spans a range of statistical and deep learning approaches. [7] employed both traditional time series models – including AutoRegressive (AR), Moving Average (MA), and ARIMA – as well as recurrent neural networks like LSTM and GRU to forecast pulse, oxygen saturation, and blood pressure. [25] proposed a generative boosting framework that leverages a generative LSTM to first generate future sequences and then uses these synthetic steps to enhance the prediction of heart rate and systolic blood pressure by the predictive LSTM up to 20 minutes ahead. [8] introduced a graph neural network architecture tuned via reinforcement learning and Bayesian optimisation to forecast heart rate up to one hour into the future. More recently, [9] integrated Transformer-based sequence modelling with diffusion probabilistic methods to forecast heart rate, systolic, and diastolic blood pressure, demonstrating strong performance on physiological time series.

While vital sign forecasting has been widely studied, few papers address the generation of prediction intervals. Notably, [26] and [27] utilize the Temporal Fusion Transformer to forecast vital signs in ICU patients and generate prediction intervals via quantile regression. While quantile regression can be applied to any model architecture, it offers no formal coverage guarantees and tends to perform poorly in out-of-distribution settings. Despite these limitations, existing work has not explored alternative approaches – such as deriving prediction intervals from distributional uncertainty estimates with CP – for vital sign forecasting.

b) Uncertainty Estimation: Uncertainty estimates measure how certain a model is about its prediction for each instance.

Definition 1 (Uncertainty Estimation): Given an instance $\mathbf{x} \in \mathbb{R}^I$ and a trained prediction model f, an uncertainty estimation method $\sigma(\mathbf{x}; f) : \mathbb{R}^I \to \mathbb{R}$ quantifies the model's uncertainty about its prediction on the instance.

An established uncertainty estimation model is the Gaussian Process Regressor (GPR) [28], which models data as a Gaussian process – a distribution over functions – and provides uncertainty estimates through its covariance function. Recent developments include Sparse GPR [29], which improves computational efficiency. A more scalable approach, the Bayesian Neural Network (BNN) [30] models weights as probability distributions, reflecting uncertainty. However, they are computationally expensive, difficult to train, and require significant architectural modifications, which can affect task performance [31].

Deep Ensemble and MCD approximate Bayesian inference in BNNs without altering neural architectures. Both estimate uncertainty by sampling multiple predictions and computing their variance. In Deep Ensemble [12], predictions are derived from individual models in the ensemble, while MCD

[11] generates predictions from a single model with dropout activated during inference. Deep Ensembles typically provide more reliable uncertainty estimates [32] due to reduced prediction correlations but require longer training times. In contrast, MCD's multiple forward passes increase inference time, limiting its use in real-time applications. Infer-Noise [33] (IN), like MCD, estimates uncertainty by injecting Gaussian noise into intermediate layers during inference and computing the variance of predictions. However, it also requires multiple forward passes, increasing inference time. Deep evidential learning (DEL) models [34], [35] are a class of single-pass uncertainty estimation methods. Instead of predicting the target directly, DEL models predict the parameters of a distribution, from which both a prediction and an uncertainty estimate are derived. For regression, [36] predicts the parameters of a Normal Inverse-Gamma distribution, using the expectation of variance and the variance of the mean for uncertainty.

Deterministic Uncertainty Methods (DUMs) [37], [38], including RUE, differ from DEL models as they do not modify the training objective. DUMs estimate uncertainty based on an instance's distance to the training set in the model's latent space, measured relative to neighboring instances of different classes [39] or class centroids [40], [41]. RUE estimates the distance to the training set using reconstruction error; instances dissimilar to the training set are poorly reconstructed, resulting in higher errors.

c) Conformal Prediction (CP): CP [13] generates PIs with coverage $1-\alpha$ from any prediction model. This paper focuses on split CP [42], which uses a fixed calibration set for computational efficiency. Split CP involves three main steps:

1) Computes a conformal score $s\left(\mathbf{x},\mathbf{y}\right)$ for each instance in the calibration set, typically quantifying the model's error (e.g., $s\left(\mathbf{x},\mathbf{y}\right)=|\mathbf{y}-\hat{\mathbf{y}}|$; $\mathbf{y}$ and $\hat{\mathbf{y}}$ refer to the true and predicted targets, respectively).
2) Calculates $\hat{q}$, the $1-\alpha$ adjusted quantile of conformal scores within the calibration set.
3) PI for any new instance is derived as $[\hat{\mathbf{y}}-\hat{q},\ \hat{\mathbf{y}}+\hat{q}]$.

However, this approach results in a PI that is uniform in width, regardless of the specific input. The state-of-the-art approach, normalized CP, generates PIs conditioned on one-dimensional uncertainty estimates, $u(\mathbf{x})$ [14], [43]. Normalized CP achieves this by modifying the score function $s\left(\mathbf{x},\mathbf{y}\right)$; rather than using only prediction errors, normalized CP divides the errors by uncertainty:

$$s\left(\mathbf{x},\mathbf{y}\right)=\frac{|\mathbf{y}-\hat{\mathbf{y}}|}{u(\mathbf{x})} \tag{2}$$

Normalized CP derives the PI as $[\hat{y}-u(\mathbf{x})\hat{q},\ \hat{y}+u(\mathbf{x})\ \hat{q}]$, conditioning the PI on the uncertainty of each instance. Note that normalized CP allows one to use any uncertainty estimate to form the PIs (e.g., conformal RUE). These methods differ fundamentally from our approaches, which compute the PIs directly without using Equation 2.

Multi-dimensional conformal prediction methods aim to compute PIs based on multi-dimensional conformity scores. Most approaches reduce these multi-dimensional scores to a single dimension. For example, [44] uses a weighted sum of conformity scores. [45] applies optimal transport to map conformity score vectors to a uniform ball distribution, using the norm as a one-dimensional conformity score. [46] employs the Mahalanobis distance of the multi-dimensional conformity score as a scalar conformity measure. [47] clusters the conformity score space and computes scalar conformity scores within each cluster. However, these methods produce PIs with uniform width across inputs and are not input-adaptive, unlike our proposed method. Only [48] addresses this limitation by using conditional normalizing flows to map multi-dimensional conformity scores into a Gaussian latent space, where the norm serves as a one-dimensional conformity score. Nevertheless, none of these methods leverage uncertainty estimates in PI construction nor apply their approaches to healthcare datasets, as is done in our proposed methods.

III. Methodology

a) Reconstruction Uncertainty Estimate: [18], [19] introduces RUE, a distributional uncertainty estimate calculated in a single pass without modifying model architecture or training objective.

Consider a model $f_{\phi,\psi}$, trained with input $\mathbf{x}\in\mathbb{R}^I$ and output $\mathbf{y}\in\mathbb{R}^O$. The model is parameterized by ϕ and ψ, which denote the feature extractor f_ϕ and prediction head f_ψ. In a regression problem, f_ϕ outputs a feature vector of size k, which is used by f_ψ to compute continuous output values $\hat{\mathbf{y}}$, i.e., $f_{\phi,\psi}:\mathbb{R}^I\mapsto\mathbb{R}^O$. The model $f_{\phi,\psi}$ is typically trained using gradient descent to minimize the discrepancy between the true $\mathbf{y}$ and predicted $\hat{\mathbf{y}}=(f_\psi\circ f_\phi)(\mathbf{x})$ target:

$$f_{\phi,\psi}=\arg\min_{\phi,\psi}||\mathbf{y}-(f_\psi\circ f_\phi)(\mathbf{x})||^2. \tag{3}$$

To compute RUE, another model — referred to as the decoder and denoted as $g:\mathbb{R}^W\mapsto\mathbb{R}^I$ — is trained to reconstruct the input $\mathbf{x}$ of the prediction model from the latent representation produced by its feature extractor f_ϕ, which has W elements. g is trained using gradient descent to minimize the discrepancy between $\mathbf{x}$ and $\hat{\mathbf{x}}=(g\circ f_\phi)(\mathbf{x})$:

$$g=\arg\min_{g}||\mathbf{x}-(g\circ f_\phi)(\mathbf{x})||^2. \tag{4}$$

f_ϕ and g form an autoencoder with one distinction: instead of training the encoder and decoder simultaneously, RUE adopts a two-step process to avoid compromising the prediction model's performance. First, the encoder f_ϕ is trained as part of the prediction model (Equation 3). Once f_ϕ is fully trained, the decoder g is trained separately (Equation 4).

Definition 2 (Reconstruction Uncertainty Estimate): The reconstruction uncertainty estimate (RUE) of instance $\mathbf{x}$ is the difference between the actual and reconstructed input [1]:

$$\sigma(\mathbf{x};f):=\left|\mathbf{x}-(g\circ f_\phi)(\mathbf{x})\right|_1.$$

[1] $|v|_1$ is the L1-norm of vector v.

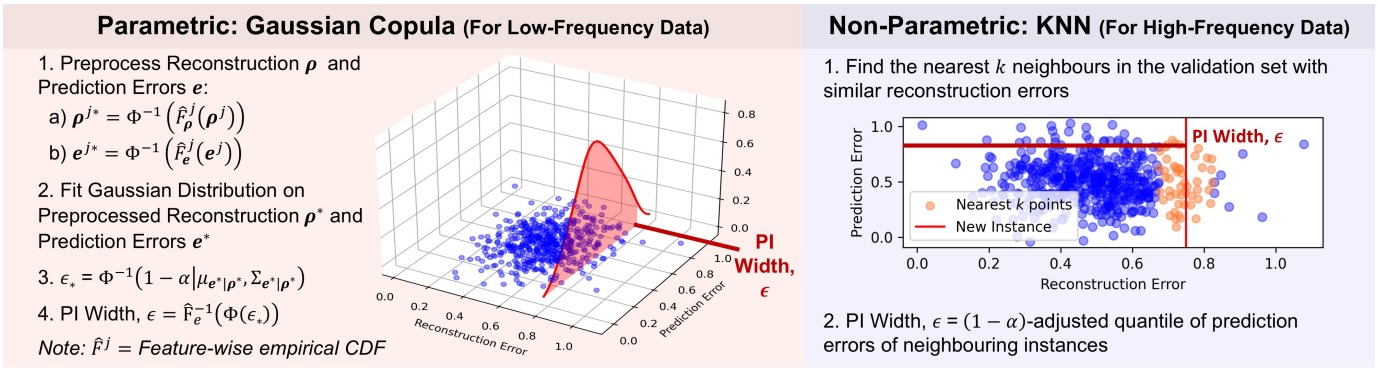


Fig. 1: Summary of proposed prediction interval (PI) methods. For the parametric Gaussian Copula PI, reconstruction and prediction errors are first transformed into a Gaussian space using their empirical CDFs $\hat{F}$ and the inverse Gaussian CDF Φ^{-1}. A joint Gaussian is then fitted to the transformed errors, and the PI width is computed from the $(1-\alpha)$-quantile of the conditional distribution of prediction errors given reconstruction errors, followed by mapping back to the original space using the inverse transforms. For the KNN PI, the PI width is computed as the $(1-\alpha)$-adjusted quantile of the prediction errors of the instance's k nearest neighbors with similar reconstruction errors.

RUE is hypothesized to reliably estimate prediction error, showing a positive correlation with prediction error:

$$|\mathbf{y} - f_{\phi,\psi}(\mathbf{x})|_1 \leftrightarrow \sigma(\mathbf{x}; f). \tag{5}$$

The above hypothesis is grounded in two assumptions:

1) The model $f_{\phi,\psi}$ performs well on inputs similar to the training set but poorly on unfamiliar inputs.
2) The decoder g, trained on the same set, reconstructs familiar inputs well, yielding low reconstruction errors (small σ), but struggles with unfamiliar inputs, resulting in high errors (large σ).

b) RUE-derived Prediction Intervals: We have devised two methods to estimate PIs from feature-wise reconstruction $\boldsymbol{\rho}$ and prediction errors $\mathbf{e}$ (Figure 1).

Definition 3 (Feature-wise Reconstruction Error): Given an instance $\mathbf{x} = [x^1, \ldots, x^I]$ and its reconstruction $\hat{\mathbf{x}} = [\hat{x}^1, \ldots, \hat{x}^I]$, its feature-wise reconstruction error is:

$$\boldsymbol{\rho} = \left[\left|x^1 - \hat{x}^1\right|, \ldots, \left|x^I - \hat{x}^I\right|\right]$$

Definition 4 (Output-wise Prediction Error): Given an instance $\mathbf{x}$, its target $\mathbf{y} = [y^1, \ldots, y^O]$ and the prediction $\hat{\mathbf{y}} = [\hat{y}^1, \ldots, \hat{y}^O]$, its output-wise prediction error is:

$$\mathbf{e} = \left[\left|y^1 - \hat{y}^1\right|, \ldots, \left|y^O - \hat{y}^O\right|\right]$$

Assuming a symmetric PI around the predicted target $\hat{\mathbf{y}}$:

$$L_i = \hat{\mathbf{y}} - \epsilon, \; U_i = \hat{\mathbf{y}} + \epsilon \tag{6}$$

We can modify Equation 1, simplifying the PI generation problem to the task of estimating the conditional distribution of the prediction error $\mathbf{e}$ given the reconstruction error $\boldsymbol{\rho}$.

$$\Pr\left(\mathbf{e}_i \leq \epsilon \mid \boldsymbol{\rho}_i\right) = 1 - \alpha \quad \equiv \quad F_{\mathbf{e}_i|\boldsymbol{\rho}_i}(\epsilon) = 1 - \alpha \tag{7}$$

Hence, each of the proposed methods aims to estimate the conditional distribution of prediction error given reconstruction error (Equation 7) and can be viewed as an extension of CP conditioned on feature-wise reconstruction error.

c) Gaussian Copula PI: The Gaussian Copula PI models the marginal distribution of each error using its empirical distribution and estimates the conditional distribution in Equation 7 with a Gaussian copula [49]. Using a Gaussian copula is motivated by the tendency of low-frequency health signals to approximate a Gaussian distribution, making the copula well-suited to model their dependencies effectively.

In practice, the Gaussian Copula PI can be interpreted as modelling the reconstruction errors $\boldsymbol{\rho}$ and prediction errors $\mathbf{e}$ using a Gaussian distribution, with additional pre- and post-processing steps. The reconstruction $\boldsymbol{\rho}$ and prediction errors $\mathbf{e}$ are preprocessed using output-wise empirical CDFs $\hat{F}$ and the inverse CDF of a Gaussian distribution, Φ^{-1}:

$$\boldsymbol{\rho}^{j*} = \Phi^{-1}\left(\hat{F}^j_{\boldsymbol{\rho}}\left(\boldsymbol{\rho}^j\right)\right), \; \mathbf{e}^{j*} = \Phi^{-1}\left(\hat{F}^j_{\mathbf{e}}\left(\mathbf{e}^j\right)\right) \tag{8}$$

After preprocessing, we fit a multivariate Gaussian distribution to the preprocessed reconstruction $\boldsymbol{\rho}^*$ and prediction errors $\mathbf{e}^*$ and compute the marginal PI as the $(1-\alpha)$-quantile of the conditional Gaussian distribution of preprocessed prediction errors $\mathbf{e}^*$ given preprocessed reconstruction errors $\boldsymbol{\rho}^*$.

$$\epsilon_* = \Phi^{-1}\left(1 - \alpha \mid \boldsymbol{\mu}_{\mathbf{e}^*|\boldsymbol{\rho}^*}, \boldsymbol{\Sigma}_{\mathbf{e}^*|\boldsymbol{\rho}^*}\right) \tag{9}$$

To derive the PI from the marginal PI, ϵ_*, we apply the following post-processing step:

$$\epsilon_{\text{copula}} = \hat{F}^{-1}_{\mathbf{e}}\left(\Phi\left(\epsilon_*\right)\right) \tag{10}$$

$\hat{F}^{-1}_{\mathbf{e}}$ represents the empirical quantile function for prediction errors $\mathbf{e}$, and Φ is the CDF of a Gaussian distribution.

Compared to CP, the Gaussian copula PI replaces the discrete score function of CP with the CDF of a Gaussian distribution for the preprocessed prediction errors $\mathbf{e}^*$ conditioned on the preprocessed reconstruction errors $\boldsymbol{\rho}^*$:

$$s_{\text{copula}}\left(\mathbf{x}, \mathbf{y}\right) = \Phi(\mathbf{e}^* \mid \boldsymbol{\mu}_{\mathbf{e}^*|\boldsymbol{\rho}^*}, \boldsymbol{\Sigma}_{\mathbf{e}^*|\boldsymbol{\rho}^*}) \tag{11}$$

Additionally, instead of applying a multiplicative correction factor to $\hat{q}$, we transform $\hat{q}$ using the CDF of a Gaussian and the inverse of the empirical CDF $\hat{F}_{\mathbf{e}}^{-1}$ (Equation 10).

d) K-Nearest Neighbours PI: The K-Nearest Neighbors (KNN) PI is a non-parametric method that estimates the conditional distribution in Equation 7 empirically. KNN is ideal for high-frequency signals because it can capture complex local patterns in the data without assuming a specific global distribution, allowing flexible modeling of rapid physiological variations. It computes the PI for a given instance, $\mathbf{x}$, by using the prediction errors, $\mathbf{e}$, of its neighbours in the reconstruction error $\boldsymbol{\rho}$-space. This is achieved through the following procedure:

1) Find the nearest $k = \lfloor\sqrt{n_v}\rceil$ neighbours in the validation set (of size n_v) with similar $\boldsymbol{\rho}$ to $\mathbf{x}$.
2) Compute the prediction errors $\mathbf{e}$ of the k neighbours.
3) The PI of $\mathbf{x}$ is the $(1-\alpha)$-adjusted quantile ($\lceil(k+1)*(1-\alpha)\rceil/k$) of prediction errors $\mathbf{e}$ of its neighbours.

The KNN PI method estimates the distribution of PIs conditioned on the reconstruction error through Steps 1 and 2. Step 3 calculates the PI with $1-\alpha$ coverage from this estimated distribution. For smaller validation sets ($n_v < \left(\frac{2}{\alpha}-1\right)^2$), using the rule-of-thumb setting $k = \lfloor\sqrt{n_v}\rceil$ may lead to poor PI generation, as the adjusted quantile causes the interval to rely on the maximum neighboring prediction errors – making it highly susceptible to outliers in the validation set. Instead, choosing a k greater than or equal to $(\frac{2}{\alpha}-1)$ (e.g., 80 in our experiments on PhysioNet) yields better PIs. We include ablation studies in the code repository, varying the KNN parameter k to assess its impact on prediction interval performance.

Viewed through the lens of CP, the KNN PI uses the same score function as standard CP but limits the calibration set to instances with similar reconstruction errors, conditioning the PI on uncertainty.

IV. Experiment

A. Data Sets

To evaluate the effectiveness of RUE-derived PI methods for vital-sign prediction models, we applied them to two large, publicly available datasets: MIMIC and PhysioNet. MIMIC contains minute-level vital signs capturing detailed physiological changes, while PhysioNet provides hour-level data that reflects longer-term trends. This difference in temporal resolution (Figure 2) allows us to test PI methods across varying signal dynamics. Additionally, the datasets differ in calibration set size – PhysioNet has a smaller calibration set – enabling us to examine the impact of limited calibration data on PI performance.

*1) **MIMIC**:* The MIMIC dataset [50] comprises vital signs from 121 intensive care unit (ICU) patients, each with a potentially different set of signals. After preprocessing (see Appendix A), the dataset consists of 59,566 instances, with 38,769, 8,765, and 12,032 instances in the training, validation, and test sets, respectively. With the preprocessed data, we trained the prediction model, denoted as $f_{\phi,\psi}(\mathbf{x})$, to forecast all six patient's states for one, two and three minutes into the future (y_{t+1}, y_{t+2}, y_{t+3}), utilizing the patient's signals from the past 5 minutes (x_{t-5} to x_t).

*2) **PhysioNet**:* The PhysioNet Challenge 2012 dataset [51] comprises signals from 2,485 intensive care unit (ICU) patients. We selected 5 vital signs from the 37 available, as they had the fewest missing values: Diastolic ABP (Arterial Blood Pressure) (mmHg), Mean ABP (mmHg), Systolic ABP (mmHg), Heart Rate (bpm), and Urine Output (mL). Set A and Set B were used for training and testing, respectively, following the Challenge's specifications. After preprocessing, the dataset consisted of 25,105 instances, with 11,608, 1,244, and 12,253 instances in the training, validation, and test sets, respectively. We trained the prediction model to forecast all five patient's states for one, two and three hours into the future (y_{t+1}, y_{t+2}, y_{t+3}), utilizing the patient's signals from the past 5 hours (x_{t-5} to x_t).

B. Baseline Models

In our experiments, we set $\alpha = 0.05$ and compare our proposed PIs against the following baselines:

*a) **RUE CP**:* We compare the multidimensional RUE methods with aggregated single-dimension RUE combined with normalized conformal prediction (CP) [14], to demonstrate the benefits of conditioning on multidimensional RUE.

Additionally, we compare our PIs with those derived by applying normalized CP [14] to five other uncertainty estimates. We selected single-model uncertainty estimates – comparable to RUE – that capture different types of uncertainty.

*b) **Monte Carlo Dropout (MCD) CP**:* We applied a 0.20 dropout rate to the prediction model's penultimate layer and generated 10 predictions per input, following hyperparameters used in [11].

*c) **Sparse Gaussian Process Regressor (SGPR) CP**:* We used *GPFlow*'s implementation of SGPR [29], [52] with a radial basis function kernel, selecting 39 and 103 random inducing variables (0.1% and 1% of training data) for MIMIC and PhysioNet respectively.

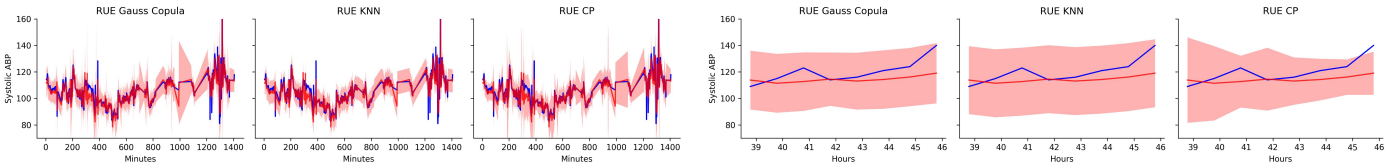
*d) **Infer-Noise (IN) CP**:* We select $\sigma \in [0.00001, 0.5]$ [33] via grid search to minimize mean absolute prediction error and sample 10 predictions per input during inference.

*e) **Bayesian Neural Network (BNN) CP**:* : We implemented BNNs [30] using *torchbnn* [53] with Bayesian linear layers (prior mean: 0, standard deviation: 0.1) and sampled 10 predictions per input.

*f) **Deep Evidential Regression (DER) CP**:* : We implemented DER using *TorchUncertainty* [54] with regularization weight = 0.01, following [36].

C. Evaluation Metrics

We assess PI quality using three metrics derived from the PI cost function in [55]:



(a) High-Frequency Data: MIMIC

(b) Low-Frequency Data: PhysioNet

Fig. 2: Line plots showing the prediction intervals (light red shading), model predictions (dark red line), and true signal (blue line) for each prediction interval method at time horizon $t+1$. For MIMIC, we present the intervals for patient "465n" on the "ABPsys (mmHg)" feature. For PhysioNet, we show the intervals for patient "142692" on the "SysABP" feature.

TABLE I: Brief descriptions of each prediction interval method, the calibration set and conformal score used, how the prediction width is computed, and whether the method is parametric or non-parametric. "1-D UE" indicates whether the method can be conditioned on one-dimensional uncertainty estimates, while "Multi-D UE" indicates whether the method can be conditioned on multi-dimensional uncertainty estimates. We indicate our proposed methods with underlining.

Prediction Interval	**Description**	**Calibration Set**	**Conformal Score**	**Prediction Width**	**Parametric**	**1-D UE**	**Multi-D UE**
Gauss Copula	Fit Gaussian Copula on uncertainty scores and prediction errors.	Whole	$\Phi(\mathbf{e}^* \mid \boldsymbol{\mu}_{\mathbf{e}^*\mid\boldsymbol{\rho}^*}, \boldsymbol{\Sigma}_{\mathbf{e}^*\mid\boldsymbol{\rho}^*})$	$\hat{F}^{-1}_{\mathbf{e}}(\Phi(\hat{q}))$	Yes	Yes	Yes
KNN	CP with a calibration set of instances with similar uncertainty.	k instances with similar uncertainty	$\lvert\mathbf{y}-\hat{\mathbf{y}}\rvert$	$\hat{q}_a = (1-\alpha)$-th adjusted quantile of scores	No	Yes	Yes
Conformal Prediction (CP)	PI widths are estimated as the adjusted quantile of calibration set scores.	Whole		$\hat{q}_a$	No	No	No
Normalised CP	Extends CP by normalizing the conformal score with uncertainty.		$\frac{\lvert\mathbf{y}-\hat{\mathbf{y}}\rvert}{u(\mathbf{x})}$	$\hat{q}_a \times u(\mathbf{x})$	No	Yes	No

*a) **Coverage Penalty (CovP):*** CovP measures the deviation of Prediction Interval Coverage Probability (PICP) from the ideal $1-\alpha$ coverage, with lower CovP indicating a better PI. PICP is the proportion of points within the PI.

$$\text{CovP} = (1-\alpha+\delta-\text{PICP})^2, \ \text{PICP} = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{y_i \in [L_i, U_i]} \quad (12)$$

n represents the total number of points, $\mathbf{y}_i$ denotes the ith ground-truth output, L_i and U_i stand for the lower and upper bounds of the PI and $\delta = \alpha/50$ is the coverage margin.

*b) **Prediction Interval Normalized Average Width (PINAW):*** The average normalized size of the PI. When two PIs have the same CP, the PI with the smaller PINAW is better.

$$\text{PINAW} = \frac{1}{n \times R}\sum_{i=1}^{n}(U_i - L_i) \quad (13)$$

R is the range of the output variable.

*c) **Coverage Width Failure Distance Criteria (CWFDC):*** It combines the two metrics above with Prediction Interval Normalized Average Failure Distance (PINAFD) to quantify the overall quality of the prediction interval (PI); a lower CWFDC indicates a better PI.

$$\text{CWFDC} = \text{PINAW} + \rho \cdot \text{PINAFD} + \beta \cdot \text{CovP} \quad (14)$$

Here, we set $\rho = 1$ and $\beta = 1000$, following the parameter values specified in [55].

PINAFD is the average normalized distance of points outside the PI. A smaller PINAFD is preferable, as it indicates that outliers are closer to the PI.

$$\text{PINAFD} = \frac{\sum_{i=1}^{n} \mathbb{1}_{y_i \notin [L_i, U_i]} \min(|y_i - U_i|, |L_i - y_i|)}{R \times \sum_{i=1}^{n}(\mathbb{1}_{y_i \notin [L_i, U_i]})} \quad (15)$$

PINAFD is zero in the absence of outliers.

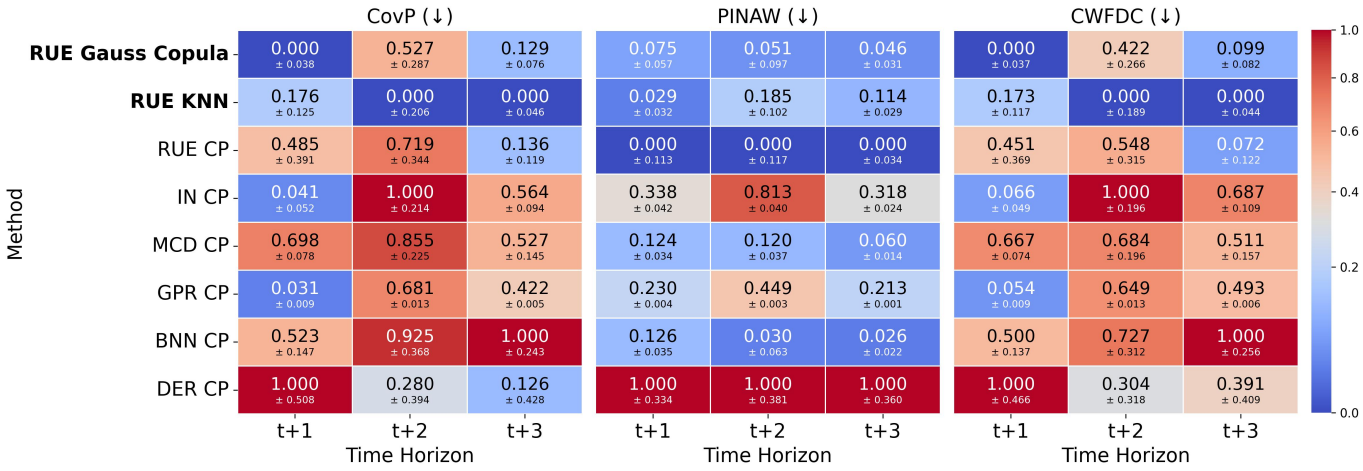
D. Experiment Results

Figure 3 compares all PIs on the MIMIC and PhysioNet datasets in terms of coverage penalty (CovP), PI width (PINAW), and overall PI quality (CWFDC).

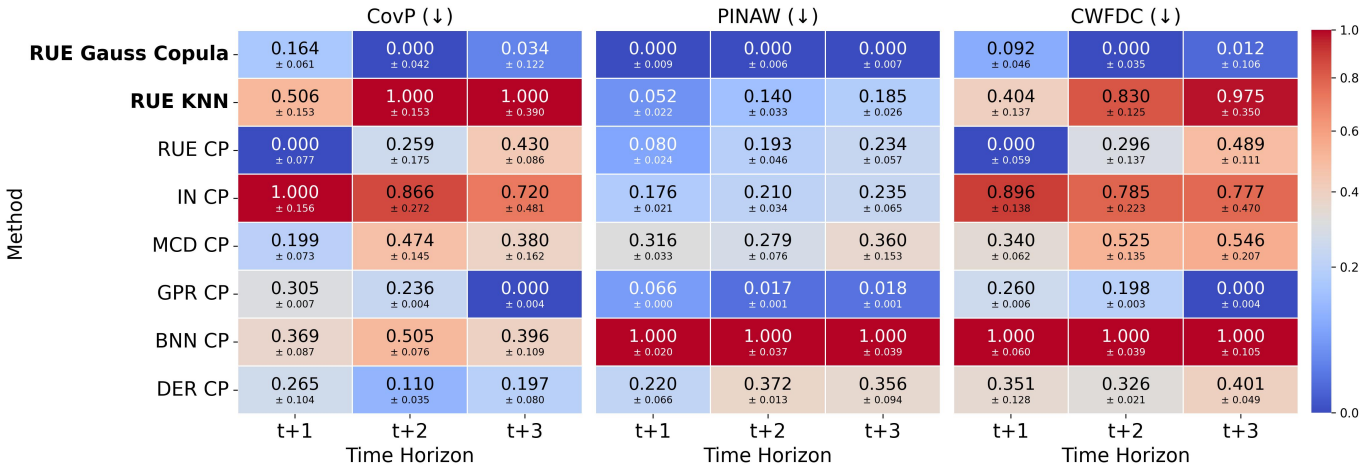
*a) **MIMIC:*** KNN and Gaussian Copula PIs yield the best overall performance, achieving the lowest CWFDC among all methods. Examining the components of CWFDC, we find that both methods achieve the best coverage – surpassing even RUE with CP.

These observations highlight the benefit of leveraging feature-wise rather than aggregated uncertainty, as is done in CP. Aggregated uncertainty reduces the likelihood of detecting anomalies in individual features – for example, peaks in heart rate – which can be hidden when averaged across all features. As a tradeoff for improved coverage, both methods produce slightly wider PIs (wider intervals increase coverage).

*b) **PhysioNet:*** Comparing the PIs based on CWFDC, the Gaussian Copula PI emerges as the best performer, followed closely by RUE with CP. When examining each component of CWFDC, we again observe Gaussian Copula's superior coverage compared to RUE CP. This is also evident in the



(a) MIMIC



(b) PhysioNet

Fig. 3: Heatmaps comparing prediction intervals in terms of Coverage Penalty (CovP; ↓), Prediction Interval Normalized Average Width (PINAW; ↓) and Coverage Width Failure Distance Criteria (CWFDC; ↓) on MIMIC and PhysioNet. For easier analysis, we min-max normalise each metric within each time horizon. The proposed PIs are highlighted in **bold text**. CovP measures the deviation of the proportion of points within the PI from the ideal coverage of 0.95, while PINAW measures PI size. CWFDC measures the overall quality of the PIs. Since all three metrics are penalties, lower values (in blue) are preferable. Each row represents a PI's performance; an ideal row would be entirely deep blue. However, trade-offs between the metrics make this unlikely. Comparing CWFDCs, the Gaussian Copula PI is the best PI, achieving the lowest CovP across datasets without compromising PINAW.

qualitative evaluations shown in Figure 2b: RUE CP's prediction interval failed to capture the drop in Systolic ABP at $t = 46$, unlike the other multidimensional RUE-based PIs. This further highlights the advantage of leveraging feature-wise uncertainty. In addition to its improved coverage, Gaussian Copula also achieves the smallest interval width. By contrast, KNN PI exhibited poorer coverage on this dataset, a deviation from the trend observed on MIMIC.

*c) **Gaussian Copula vs. KNN**:* The choice between Gaussian Copula and KNN for PI generation depends on three key dataset attributes: (1) the size of the calibration set, (2) the distribution of errors, and (3) the temporal resolution of the data. In our experiments, we found that when the calibration set is small (as in PhysioNet), KNN struggles to estimate PIs accurately, resulting in wider intervals. Gaussian Copula tends to perform better when the reconstruction and prediction errors are closer to Gaussian, as indicated by the Henze–Zirkler (HZ) test statistic. For example, the HZ statistic is lower for PhysioNet (1.5) than for MIMIC (2.4), suggesting that the errors in PhysioNet are more Gaussian – consistent with our observation that Gaussian Copula performs better on PhysioNet than on MIMIC. Furthermore, Gaussian Copula performed the worst on MIMIC at t+2, where the HZ statistic was particularly high (3.3). Temporal resolution also plays a

role: minute-level data in MIMIC captures local fluctuations, allowing KNN to find similar short-term patterns and produce intervals with better coverage, whereas the hour-level data in PhysioNet is smoother and more Gaussian-like, favoring the parametric approach of Gaussian Copula.

V. Discussion and Conclusion

This paper introduces two methods – Gaussian Copula and KNN – for deriving prediction intervals (PIs) from the multidimensional Reconstruction Uncertainty Estimate. These methods focus on estimating the conditional distribution of the PI based on the input's uncertainty. The Gaussian Copula models this distribution directly by fitting a multivariate Gaussian Copula to reconstruction and prediction errors. The KNN method estimates the PI empirically using prediction errors of instances with similar reconstruction errors. Experiments on the MIMIC and PhysioNet datasets show that the Gaussian Copula method performs consistently well across both datasets, achieving high coverage with competitive widths. The KNN method outperforms on MIMIC, where the higher temporal resolution and larger calibration set favour its local estimation approach.

As all proposed PIs support multidimensional uncertainty estimates, future studies could explore using them to ensemble uncertainty estimates and capture all sources of prediction uncertainty (aleatoric, epistemic, and distributional). Additionally, while RUE-based PIs show promise, like CP, their accuracy depends on validation set size, with smaller sets affecting reliability. Future work could develop direct PI generation methods, such as training an uncertainty calibration network with a loss function optimizing coverage and interval width, reducing reliance on validation sets.

References

[1] W. Q. Mok, W. Wang, and S. Y. Liaw, "Vital signs monitoring to detect patient deterioration: An integrative literature review," *International journal of nursing practice*, vol. 21, pp. 91–98, 2015.

[2] M. Cardona-Morrell, M. Nicholson, and K. Hillman, "Vital signs: From monitoring to prevention of deterioration in general wards," in *Annual Update in Intensive Care and Emergency Medicine 2015*. Springer, 2015, pp. 533–545.

[3] I. J. Brekke, L. H. Puntervoll, P. B. Pedersen, J. Kellett, and M. Brabrand, "The value of vital sign trends in predicting and monitoring clinical deterioration: A systematic review," *PloS one*, vol. 14, no. 1, p. e0210875, 2019.

[4] T. Kenzaka, M. Okayama, S. Kuroki, M. Fukui, S. Yahata, H. Hayashi, A. Kitao, D. Sugiyama, E. Kajii, and M. Hashimoto, "Importance of vital signs to the early diagnosis and severity of sepsis: association between vital signs and sequential organ failure assessment score in patients with sepsis," *Internal Medicine*, vol. 51, no. 8, pp. 871–876, 2012.

[5] Y. Eddahchouri, R. V. Peelen, M. Koeneman, H. R. Touw, H. van Goor, and S. J. Bredie, "Effect of continuous wireless vital sign monitoring on unplanned icu admissions and rapid response team calls: a before-and-after study," *British Journal of Anaesthesia*, vol. 128, no. 5, pp. 857–863, 2022.

[6] F. L. Becking-Verhaar, R. P. Verweij, M. de Vries, H. Vermeulen, H. van Goor, and G. J. Huisman-de Waal, "Continuous vital signs monitoring with a wireless device on a general ward: a survey to explore nurses' experiences in a post-implementation period," *International Journal of Environmental Research and Public Health*, vol. 20, no. 10, p. 5794, 2023.

[7] T. Bhavani, P. VamseeKrishna, C. Chakraborty, and P. Dwivedi, "Stress classification and vital signs forecasting for iot-health monitoring," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 2022.

[8] T. Shaik, X. Tao, H. Xie, L. Li, J. Yong, and Y. Li, "Graph-enabled reinforcement learning for time series forecasting with adaptive intelligence," *IEEE Transactions on Emerging Topics in Computational Intelligence*, 2024.

[9] P. Chang, H. Li, S. F. Quan, S. Lu, S.-F. Wung, J. Roveda, and A. Li, "A Transformer-based Diffusion Probabilistic Model for Heart Rate and Blood Pressure Forecasting in Intensive Care Unit," *Computer Methods and Programs in Biomedicine*, vol. 246, p. 108060, 2024. [Online]. Available: http://arxiv.org/abs/2301.06625

[10] A. A. Abdullah, M. M. Hassan, and Y. T. Mustafa, "A Review on Bayesian Deep Learning in Healthcare: Applications and Challenges," *IEEE Access*, vol. 10, pp. 36 538–36 562, 2022.

[11] Y. Gal and Z. Ghahramani, "Dropout as a bayesian approximation: Representing model uncertainty in deep learning," in *ICML*. PMLR, 2016, pp. 1050–1059.

[12] B. Lakshminarayanan, A. Pritzel, and C. Blundell, "Simple and Scalable Predictive Uncertainty Estimation using Deep Ensembles," in *Advances in Neural Information Processing Systems*, vol. 30. Curran Associates, Inc., 2017.

[13] A. Gammerman, V. Vovk, and V. Vapnik, "Learning by transduction, vol uai'98," 1998.

[14] H. Papadopoulos, A. Gammerman, and V. Vovk, "Normalized nonconformity measures for regression conformal prediction," in *Proceedings of the IASTED AIA 2008*, 2008, pp. 64–69.

[15] Y. Renkema, N. Brinkel, and T. Alskaif, "Conformal Prediction for Stochastic Decision-Making of PV Power in Electricity Markets," 2024. [Online]. Available: http://arxiv.org/abs/2403.20149

[16] I. D. Khurjekar and P. Gerstoft, "Uncertainty quantification for direction-of-arrival estimation with conformal prediction," *The Journal of the Acoustical Society of America*, vol. 154, no. 2, pp. 979–990, 2023. [Online]. Available: https://doi.org/10.1121/10.0020655

[17] A. Timans, C.-N. Straehle, K. Sakmann, and E. Nalisnick, "Adaptive Bounding Box Uncertainties via Two-Step Conformal Prediction," 2024. [Online]. Available: http://arxiv.org/abs/2403.07263

[18] L. R. Wang, T. C. Henderson, and X. Fan, "An uncertainty estimation model for algorithmic trading agent," in *Intelligent Autonomous Systems 18*. Cham: Springer Nature Switzerland, 2024, pp. 459–465.

[19] L. Korte, L. R. Wang, and X. Fan, "Confidence estimation in analyzing intravascular optical coherence tomography images with deep neural networks," in *2024 IEEE Conference on Artificial Intelligence (CAI)*. IEEE, 2024, pp. 358–364.

[20] J. L. T. Hao, L. R. Wang, C. Liu, C. Choi, S. Liu, and X. Fan, "Exploring epistemic and distributional uncertainties in algorithmic trading agents," in *2024 IEEE International Conference on Agents (ICA)*. IEEE, 2024, pp. 82–87.

[21] L. Li, L. R. Wang, H. Fu, and X. Fan, "Trading confidence: Comprehensive uncertainty estimation in algorithmic trading," in *29th Pacific Asia Conference on Information Systems, PACIS 2025, Kuala Lumpur, Malaysia, July 5-10, 2025*, 2025.

[22] J. Rahman, A. Brankovic, M. Tracy, S. Khanna *et al.*, "Exploring computational techniques in preprocessing neonatal physiological signals for detecting adverse outcomes: Scoping review," *Interactive Journal of Medical Research*, vol. 13, no. 1, p. e46946, 2024.

[23] N. Shawen, M. K. O'Brien, S. Venkatesan, L. Lonini, T. Simuni, J. L. Hamilton, R. Ghaffari, J. A. Rogers, and A. Jayaraman, "Role of data measurement characteristics in the accurate detection of parkinson's disease symptoms using wearable sensors," *Journal of neuroengineering and rehabilitation*, vol. 17, pp. 1–14, 2020.

[24] B. L. Ortiz, V. Gupta, R. Kumar, A. Jalin, X. Cao, C. Ziegenbein, A. Singhal, M. Tewari, and S. W. Choi, "Data preprocessing techniques for ai and machine learning readiness: Scoping review of wearable sensor data in cancer care," *JMIR mHealth and uHealth*, vol. 12, no. 1, p. e59587, 2024.

[25] S. Liu, J. Yao, and M. Motani, "Early prediction of vital signs using generative boosting via lstm networks," in *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE, 2019, pp. 437–444.

[26] R. Phetrittikun, K. Suvirat, T. N. Pattalung, C. Kongkamol, T. Ingviya, and S. Chaichulee, "Temporal fusion transformer for forecasting vital sign trajectories in intensive care patients," in *2021 13th Biomedical*

Engineering International Conference (BMEiCON). IEEE, 2021, pp. 1–5.

[27] R. He and J. N. Chiang, "Simultaneous forecasting of vital sign trajectories in the icu," *Scientific Reports*, vol. 15, no. 1, p. 14996, 2025.

[28] C. K. Williams and C. E. Rasmussen, *Gaussian processes for machine learning*. MIT press Cambridge, MA, 2006, vol. 2.

[29] M. Titsias, "Variational learning of inducing variables in sparse gaussian processes," in *Artificial intelligence and statistics*. PMLR, 2009, pp. 567–574.

[30] L. V. Jospin, H. Laga, F. Boussaid, W. Buntine, and M. Bennamoun, "Hands-On Bayesian Neural Networks—A Tutorial for Deep Learning Users," *IEEE Computational Intelligence Magazine*, vol. 17, no. 2, pp. 29–48, May 2022.

[31] X. Yu, G. Franchi, J. Gu, and E. Aldea, "Discretization-Induced Dirichlet Posterior for Robust Uncertainty Quantification on Regression," *Proceedings of the AAAI CAI*, vol. 38, no. 7, pp. 6835–6843, 2024. [Online]. Available: https://ojs.aaai.org/index.php/AAAI/article/view/28508

[32] N. Durasov, T. Bagautdinov, P. Baque, and P. Fua, "Masksembles for Uncertainty Estimation," in *2021 IEEE/CVF Conference on CVPR*. Nashville, TN, USA: IEEE, Jun. 2021, pp. 13 534–13 543.

[33] L. Mi, H. Wang, Y. Tian, H. He, and N. N. Shavit, "Training-Free Uncertainty Estimation for Dense Regression: Sensitivity as a Surrogate," *Proceedings of AAAI*, vol. 36, no. 9, pp. 10 042–10 050, 2022.

[34] D. Ulmer, C. Hardmeier, and J. Frellsen, "Prior and Posterior Networks: A Survey on Evidential Deep Learning Methods For Uncertainty Estimation," *Transactions on Machine Learning Research*, 2023.

[35] B. Charpentier, D. Zügner, and S. Günnemann, "Posterior Network: Uncertainty Estimation without OOD Samples via Density-Based Pseudo-Counts," in *Advances in Neural Information Processing Systems*, vol. 33. Curran Associates, Inc., 2020, pp. 1356–1367.

[36] A. Amini, W. Schwarting, A. Soleimany, and D. Rus, "Deep Evidential Regression," 2020.

[37] J. Postels, M. Segu, T. Sun, L. Sieber, L. Van Gool, F. Yu, and F. Tombari, "On the Practicality of Deterministic Epistemic Uncertainty," 2022.

[38] B. Charpentier, C. Zhang, and S. Günnemann, "Training, Architecture, and Prior for Deterministic Uncertainty Methods," 2023.

[39] A. Mandelbaum and D. Weinshall, "Distance-based Confidence Score for Neural Network Classifiers," 2017.

[40] J. V. Amersfoort, L. Smith, Y. W. Teh, and Y. Gal, "Uncertainty Estimation Using a Single Deep Deterministic Neural Network," in *Proceedings of the 37th ICML*. PMLR, 2020, pp. 9690–9700.

[41] I. Apostolopoulou, B. Eysenbach, F. Nielsen, and A. Dubrawski, "A Rate-Distortion View of Uncertainty Quantification," 2024.

[42] H. Papadopoulos, V. Vovk, and A. Gammerman, "Conformal prediction with neural networks," in *19th IEEE International ICTAI 2007*, vol. 2. IEEE, 2007, pp. 388–395.

[43] A. N. Angelopoulos and S. Bates, "A Gentle Introduction to Conformal Prediction and Distribution-Free Uncertainty Quantification," 2022. [Online]. Available: http://arxiv.org/abs/2107.07511

[44] R. Luo and Z. Zhou, "Weighted aggregation of conformity scores for classification," *arXiv preprint arXiv:2407.10230*, 2024.

[45] M. Klein, L. Bethune, E. Ndiaye, and M. Cuturi, "Multivariate conformal prediction using optimal transport," *arXiv preprint arXiv:2502.03609*, 2025.

[46] C. Xu, H. Jiang, and Y. Xie, "Conformal prediction for multi-dimensional time series by ellipsoidal sets," *arXiv preprint arXiv:2403.03850*, 2024.

[47] Y. Tawachi and B. Laufer-Goldshtein, "Multi-dimensional conformal prediction," in *The Thirteenth International Conference on Learning Representations*, 2025.

[48] J. Lee, C. Xu, and Y. Xie, "Flow-based conformal prediction for multi-dimensional time series," *arXiv preprint arXiv:2502.05709*, 2025.

[49] K. Aas, M. Jullum, and A. Løland, "Explaining individual predictions when features are dependent: More accurate approximations to Shapley values," *Artificial Intelligence*, vol. 298, p. 103502, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0004370221000539

[50] G. B. Moody and R. G. Mark, "A database to support development and evaluation of intelligent intensive care monitoring," in *Computers in Cardiology 1996*. IEEE, 1996, pp. 657–660.

[51] I. Silva, G. Moody, D. J. Scott, L. A. Celi, and R. G. Mark, "Predicting in-hospital mortality of icu patients: The physionet/computing in cardiology challenge 2012," in *2012 Computing in Cardiology*. IEEE, 2012, pp. 245–248.

[52] A. G. d. G. Matthews, M. van der Wilk, T. Nickson, K. Fujii, A. Boukouvalas, P. León-Villagrá, Z. Ghahramani, and J. Hensman, "GPflow: A Gaussian process library using TensorFlow," *Journal of Machine Learning Research*, vol. 18, no. 40, pp. 1–6, apr 2017. [Online]. Available: http://jmlr.org/papers/v18/16-537.html

[53] S. Lee, H. Kim, and J. Lee, "Graddiv: Adversarial robustness of randomized neural networks via gradient diversity regularization," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022.

[54] A. Lafag and O. Laurent, "Torchuncertainty," 2024, version 0.3.0. [Online]. Available: torch-uncertainty.github.io/

[55] H. D. Kabir, A. Khosravi, S. Nahavandi, and D. Srinivasan, "Neural network training for uncertainty quantification over time-range," *IEEE Transactions on Emerging Topics in Computational Intelligence*, vol. 5, no. 5, pp. 768–779, 2020.

[56] J. H. Friedman, J. L. Bentley, and R. A. Finkel, "An algorithm for finding best matches in logarithmic expected time," *ACM Transactions on Mathematical Software (TOMS)*, vol. 3, no. 3, pp. 209–226, 1977.

[57] Y. Ding, J. Liu, J. Xiong, and Y. Shi, "Revisiting the evaluation of uncertainty estimation and its application to explore model complexity-uncertainty trade-off," in *Proceedings of the IEEE/CVF CVPR Workshops*, 2020, pp. 4–5.

APPENDIX A
DATASET PRE-PROCESSING

Preprocessing steps for MIMIC dataset:

1) Select patients with all six vital signs: ["ABPdias (mmHg)", "ABPmean (mmHg)", "ABPsys (mmHg)", "HR (bpm)", "RESP (bpm)", "SpO2 (%)"]. A total of 57 patients were selected.
2) Randomly assign 70%, 10%, and 20% of patients to the training, validation, and test sets, respectively.
3) Remove all feature values less than 0 or blood pressure values greater than 250.
4) Z-normalize the signals using the mean and standard deviation derived from the training set.
5) To reduce the signal frequency, downsample the signal to one sample per minute by calculating the mean and standard deviation of signals within each minute.
6) Drop rows with missing values.

Preprocessing steps for PhysioNet Challenge dataset:

1) Rows with missing values were dropped.
2) Rows with anomalous values were dropped, such as when Diastolic, Mean, or Systolic ABP was 0 or below, or Urine Output exceeded 1,000 mL.
3) Data were discretized into 1-hour intervals, as this was the most common sampling frequency.
4) Set A was split into training and validation subsets on a per-patient basis; 10% of patients were randomly assigned to the validation set.
5) All signals were min-max normalized.

"ABPdias (mmHg)", "ABPmean (mmHg)", "ABPsys (mmHg)" signals correspond to the diastolic, mean, and systolic arterial blood pressure of the patient, while "HR (bpm)", "RESP (bpm)", "SpO2 (%)" correspond to periodic measurements of the heart rate, respiration rate, and oxygen saturation of the patient.

APPENDIX B
IMPLEMENTATION DETAILS

To expedite the search process in step 1 of KNN PI, we employ a K-Dimensional Tree [56] fitted on validation set reconstruction errors before inference, and queried with the reconstruction errors of $\mathbf{x}$ during inference. We applied RUE, MCD, and IN to an MLP prediction model, using an MLP decoder for RUE. For RUE, encoder and decoder width/depth were fine-tuned via grid search, with validation loss and correlation with loss as their respective metrics. Regularisation and early stopping (patience $= 20$) mitigated overfitting. Hyperparameters of other baselines were tuned similarly via grid search using validation performance. All experiments were repeated five times, reporting mean and standard deviation of all metrics. For implementation details, hyperparameter settings, and ablation studies on the KNN parameter k, refer to the code repository: https://github.com/lr98769/rue_prediction_interval.

APPENDIX C
EVALUATING TIME SERIES PREDICTION PERFORMANCE

TABLE II: Comparison of Mean Squared Error of Models Across All Time Horizons. In each column, we **bold** the values for the best-performing model and underline the values for the second-best-performing model.

Model	MIMIC	PhysioNet
RUE	0.1008 ± 0.000	0.0024 ± 0.000
MCD	0.1017 ± 0.000	0.0024 ± 0.000
GPR	**0.0962 ± 0.000**	**0.0023 ± 0.000**
IN	0.1008 ± 0.000	0.0024 ± 0.000
BNN	0.1028 ± 0.000	0.0053 ± 0.000
DER	0.1010 ± 0.000	0.0026 ± 0.000

Table II presents the average forecasting performance of various uncertainty estimation models across three time horizons. For the MIMIC dataset, SGPR achieved the best prediction performance, followed by IN and RUE (or multilayer perceptron (MLP)), as evidenced by their low MSE values across all horizons. On the PhysioNet dataset, RUE, MCD, SGPR and IN had comparable performances. Across both datasets, BNN and DER exhibited the weakest performance. These results demonstrate that the choice of uncertainty estimation method can significantly influence prediction performance.

APPENDIX D
EVALUATING UNCERTAINTY ESTIMATION PERFORMANCE

We evaluate uncertainty estimates using three metrics, with prediction loss measured by mean absolute error (MAE).

*a) **Correlation**:* The correlation between the uncertainty estimate and prediction error [33], where a strong positive correlation indicates a reliable estimate.

*b) **AURC**:* The area under the risk-coverage curve [57], which plots prediction loss (risk) against the proportion of confident predictions (coverage). Lower AURC indicates a more selective uncertainty estimate.

TABLE III: Comparison of Uncertainty Estimation Performance on MIMIC and PhysioNet Across Time Horizons. For each metric, we **bold** values for the best-performing model and underline values for the second-best-performing model.

Data	t+k	Model	Correlation (↑)	AURC (↓)	σ = 0.1 (↓)	σ = 0.2 (↓)
MIMIC	t+1	RUE	**0.431 ± 0.038**	0.095 ± 0.003	0.062 ± 0.007	0.079 ± 0.004
		MCD	0.192 ± 0.028	0.125 ± 0.001	0.103 ± 0.006	0.121 ± 0.002
		GPR	0.205 ± 0.002	0.084 ± 0.000	0.070 ± 0.003	0.076 ± 0.000
		IN	0.086 ± 0.025	0.128 ± 0.002	0.124 ± 0.009	0.126 ± 0.005
		BNN	0.285 ± 0.017	0.116 ± 0.001	0.081 ± 0.004	0.104 ± 0.004
		DER	0.242 ± 0.035	**0.074 ± 0.002**	**0.057 ± 0.003**	**0.063 ± 0.002**
	t+2	RUE	**0.384 ± 0.019**	**0.109 ± 0.005**	**0.088 ± 0.008**	**0.096 ± 0.007**
		MCD	0.132 ± 0.010	0.151 ± 0.001	0.113 ± 0.005	0.142 ± 0.008
		GPR	0.170 ± 0.003	0.112 ± 0.000	0.102 ± 0.001	0.103 ± 0.000
		IN	-0.002 ± 0.007	0.156 ± 0.002	0.173 ± 0.033	0.163 ± 0.006
		BNN	0.203 ± 0.009	0.145 ± 0.003	0.108 ± 0.011	0.132 ± 0.004
		DER	0.160 ± 0.066	0.112 ± 0.002	0.093 ± 0.003	0.100 ± 0.004
	t+3	RUE	**0.366 ± 0.012**	0.138 ± 0.006	**0.101 ± 0.024**	0.125 ± 0.006
		MCD	0.120 ± 0.023	0.172 ± 0.003	0.122 ± 0.005	0.165 ± 0.005
		GPR	0.152 ± 0.003	**0.132 ± 0.001**	0.122 ± 0.002	**0.121 ± 0.000**
		IN	-0.001 ± 0.007	0.178 ± 0.005	0.160 ± 0.035	0.173 ± 0.010
		BNN	0.174 ± 0.012	0.167 ± 0.001	0.124 ± 0.006	0.157 ± 0.003
		DER	0.120 ± 0.020	0.139 ± 0.004	0.120 ± 0.002	0.132 ± 0.009
Physio Net	t+1	RUE	**0.282 ± 0.014**	0.025 ± 0.001	**0.020 ± 0.002**	**0.022 ± 0.001**
		MCD	0.189 ± 0.012	0.026 ± 0.001	0.024 ± 0.001	0.025 ± 0.001
		GPR	0.174 ± 0.001	**0.024 ± 0.000**	0.022 ± 0.000	0.024 ± 0.000
		IN	0.057 ± 0.012	0.028 ± 0.001	0.028 ± 0.003	0.027 ± 0.001
		BNN	0.122 ± 0.004	0.050 ± 0.000	0.049 ± 0.003	0.048 ± 0.001
		DER	0.236 ± 0.013	0.025 ± 0.001	0.022 ± 0.001	0.023 ± 0.001
	t+2	RUE	**0.232 ± 0.007**	0.029 ± 0.000	**0.025 ± 0.001**	**0.026 ± 0.000**
		MCD	0.149 ± 0.007	0.030 ± 0.000	0.029 ± 0.001	0.028 ± 0.001
		GPR	0.148 ± 0.000	**0.028 ± 0.000**	0.026 ± 0.000	0.028 ± 0.000
		IN	0.035 ± 0.006	0.032 ± 0.000	0.042 ± 0.009	0.031 ± 0.002
		BNN	0.111 ± 0.005	0.051 ± 0.000	0.049 ± 0.003	0.048 ± 0.001
		DER	0.173 ± 0.007	0.029 ± 0.001	0.029 ± 0.002	0.027 ± 0.001
	t+3	RUE	**0.231 ± 0.009**	0.031 ± 0.001	**0.028 ± 0.001**	**0.028 ± 0.001**
		MCD	0.136 ± 0.015	0.033 ± 0.000	0.032 ± 0.001	0.031 ± 0.001
		GPR	0.147 ± 0.000	**0.030 ± 0.000**	0.029 ± 0.000	0.030 ± 0.000
		IN	0.030 ± 0.007	0.034 ± 0.000	0.038 ± 0.005	0.033 ± 0.002
		BNN	0.105 ± 0.007	0.051 ± 0.000	0.049 ± 0.003	0.049 ± 0.001
		DER	0.186 ± 0.019	0.032 ± 0.001	0.031 ± 0.003	0.031 ± 0.001

*c) **σ-Risk Score**:* Measures prediction error for confident predictions (normalized uncertainty $\leq \sigma \in [0.1, 0.2]$), reflecting robustness to false negatives (incorrect predictions with high confidence). Lower values indicate greater resilience. To mitigate the impacts of outliers, uncertainty is normalized by (1) excluding values beyond 1.5 IQR above Q3 and (2) applying min–max normalization on the remaining range.

Table III compares uncertainty estimate performance across datasets and horizons. Reliability generally decreases with longer horizons, as shown by the declining correlation with prediction error from $t+1$ to $t+3$, suggesting existing estimates struggle to capture uncertainty at extended horizons.

Correlation: RUE is the most reliable estimate on both MIMIC and PhysioNet, followed by BNN on MIMIC and DER on PhysioNet.

AURC: SGPR achieves the lowest AURC, followed closely by RUE, particularly beyond $t+1$. This highlights RUE's potential for selective prediction (Table II).

σ-Risk Scores: RUE is the most robust to false negatives, showing the smallest or second-smallest σ-risk across datasets and horizons, followed by DER and SGPR, indicating its confident predictions are less likely to be incorrect.